

A divided world

The lack of preparation for last month's tsunami illustrates shocking disparities in how science is applied in different regions of the world. The global response to the disaster offers a glimmer of hope that these disparities will be addressed.

As the full horror of the Asian tsunami sinks in, the reactions of scientists echo those of the population as a whole. These range from a sense of hopelessness in the face of nature's power to concern for the victims and a determination that their suffering should be addressed.

The Indian Ocean tsunami of 26 December 2004 occurred at about 01:00 GMT, when the Indian tectonic plate moved underneath the neighbouring Burma microplate, raising it by about 10 metres along a length of more than 1,000 km and sending a wave propagating through the full depth of the overlying ocean at high speed. With wavelengths much larger than the depth of the ocean, such waves propagate across the great distances of the open sea without much surface perturbation and with very little energy loss, until shallower coastal shelves slow the wave and increase its amplitude — resulting, in this case, in a calamity of biblical proportions.

Such disasters have always been with us, but this particular event (see News, pages 3–5) had some characteristics that cry out for a global response that is more emphatic and sustained than a brief outburst of charity.

The most distinctive of these characteristics is the uneasy feeling, prompted by the delayed action of the tsunami, that a great deal of the suffering could have been avoided. Much of the damage, after all, occurred in Sri Lanka and on India's eastern coast about two hours after an earthquake had triggered the tsunami in the ocean. Monitoring stations in Japan and the United States, for example, had been able to observe the event in real time and yet apparently could do nothing — despite the ubiquity of modern telecommunications — to warn victims of the impending risk.

It turns out, on closer examination, that not all of this is true. The size of the earthquake wasn't apparent at first glance: early estimates put it at magnitude 8, which is not exceptional for submarine quakes and is an order of magnitude smaller than the eventual value of 9 that made this the world's largest seismic event for 40 years. And, in the absence of an ocean-based monitoring system, remote seismologists did not know that the quake had triggered a tsunami. Many researchers who were alerted to the event in the United States on their Christmas night, for example, went to bed quite oblivious to the carnage that was unfolding as they slept.

Additionally, as the awful scale of the disaster slowly emerged from remote regions of western Indonesia, it has become clear that most of the death and destruction had occurred in a region that was too close to the epicentre of the event for warnings to have made much difference.

Neglect

Nonetheless, an effective warning system, allied to a public education campaign of the sort that has already taken place around the Pacific Ocean, could have reduced the scale of the disaster.

It is clear, with the benefit of hindsight, that the arcane international bodies that manage tsunami protection have been neglected and underfunded for many years. Most of them have focused on the Pacific Ocean, and occasional attempts to widen their brief to the Indian Ocean have been rebuffed.

A master plan prepared in 1999 by ITSU, one of the international organizations that plans for the monitoring of tsunamis, stated: "Tsunami hazards exist on both sides of the Atlantic Ocean, in the eastern Indian Ocean, and in the Mediterranean, Caribbean, and Black Seas. Efforts to establish warning centers in those areas should be encouraged."

An important reason for the previous confinement of monitoring systems to the Pacific has been the occurrence of two tsunamis in the Pacific quite recently, in 1960 and 1964. The last tsunami produced by an earthquake in the Indian Ocean is thought to have occurred back in 1833.

However, the most important differentiating factor has been the readiness of 'Pacific rim' nations such as Japan, Australia and the United States to support a cheap but potentially effective system for monitoring and for educating the public about an infrequent risk. India, Indonesia and the other nations on the Indian Ocean's rim are relatively poor countries with needs that seemed more pressing than that of planning against the remote — but nonetheless inevitable — prospect of a tsunami.

Pushing for change

A great amount could have been done at relatively little expense to plan for a tsunami, however. The most important component of such preparation is public education, so that local inhabitants are aware, for example, of the fact that a dramatic recession of the ocean is in itself a warning of an impending event. The next most important component is the construction of a simple network that will quickly convey warning information from the seismological stations to some central point (such as the Pacific Tsunami Warning Center in Hawaii) and back out again to local radio and television channels, perhaps using siren systems in regions that can afford them.

Some of this will doubtless now take place — and so it must. As earthquake-mitigation programmes in Japan and California have shown, we can avoid vast carnage in the face of major natural disruptions. Scientists have a role to play in this. Biomedical researchers have taken global initiatives to address preventable deaths from tropical diseases that might otherwise be ignored. In the same spirit, Earth scientists around the world must now press even harder for resources in rich countries to be brought to bear to confront the risks of natural disasters in poor countries.

The same communications technologies that could have helped to mitigate this disaster have, instead, brought it home relentlessly to our living rooms. The science behind the event has been busily and prominently displayed for all to see — alongside the consequences of inaction in the face of well-established risks.

Is it too much to expect that people in rich countries, when confronted with evidence on such a scale, will ask that their governments start to pay modest respect to the value of human life amongst the poor, and adjust their budgetary priorities accordingly? Scientists, at least, should argue for a strengthening of research priorities that reflect the needs not of well-protected interest groups in their own nations, but of humanity itself. ■





Past events
The history of
tsunamis offers
clear lessons
p4



Under attack
Indian government
faces outcry over
lack of warning
p5



Back to Earth
Balloon touches
down after hunt
for antimatter
p6



Blown out
Bid to grow
marijuana for
research fails
p7

Inadequate warning system left Asia at the mercy of tsunami

Emma Marris, Washington

When two tectonic plates beneath the Indian Ocean cracked past each other at 0:59 GMT on 26 December 2004, the sea floor was forced upwards by some 10 metres. This displaced in the region of a trillion tonnes of water, driving it towards southeast Asia's coastline in a long, low-amplitude wave travelling at up to 900 kilometres per hour.

When the wave reached shallower water near the coast, it shortened, slowed and gathered into surges that killed at least 150,000 people across a dozen countries. In the aftermath of the disaster, casualties continue to mount at a ferocious pace.

Seismologists knew about the magnitude 9 earthquake within minutes (see 'Triple slip of tectonic plates caused seafloor surge', below), but the absence of monitoring equipment in the ocean itself meant that they didn't know for sure that a tsunami had occurred. Those who suspected as much were unsure how to get the word out to the regions most at risk.

Although the small global community of tsunami researchers had expressed some concerns about the risk of such an event, little had been done to plan for it. "It is



Devastated: the shattered remains of Meulaboh in Indonesia, largely destroyed by the tsunami.

Triple slip of tectonic plates caused seafloor surge

In the aftermath of the tsunami that devastated coastlines around the Indian Ocean, experts are piecing together details of the seismic slip that sparked the waves. The earthquake, the world's biggest for more than 40 years and the fourth largest since 1900, has literally redrawn the map, moving some islands by up to 20 metres.

The destruction, which claimed as many as 150,000 lives, was unleashed by a 'megathrust' — a sudden juddering movement beneath the sea floor. A build-up of pressure caused the floor of the Indian Ocean to lurch some 15 metres towards Indonesia, burrowing under a tectonic plate and triggering the ferocious swells that smashed into surrounding shores.

The earthquake followed almost two centuries of tension during which the India plate pressed against the Burma microplate, which

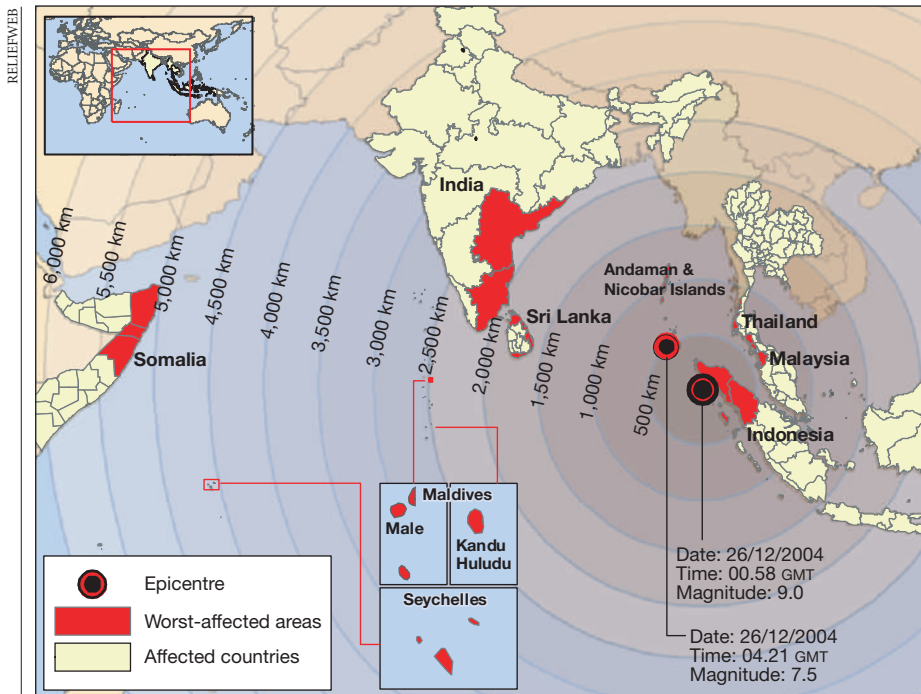
carries the tip of Sumatra as well as the Andaman and Nicobar Islands. The plates move against one another at an average rate of about 6 centimetres a year, but this movement does not occur smoothly. There has not been a very large quake along this fault since 1833 — a fact that may have contributed to the huge force of this one. The India plate's jarring slide released the tension on the Burma microplate, causing it to spring violently upwards.

Quakes of this type, called subduction earthquakes, are commonplace throughout the world, but rarely strike with such force, says Roger Musson of the British Geological Survey in Edinburgh. "This is the largest earthquake I've seen in my career as a seismologist," he says. "The length of the rupture was 1,200 kilometres — I could hardly believe it."

The earthquake, measured at magnitude 9.0, actually consisted of three events that occurred within seconds of each other, Musson explains. The initial slip, which happened to the west of Sumatra's northern tip, triggered two further slips to the north. The total force released was enough to jolt the entire planet.

The seafloor bulge unleashed a wave that surged through the Indian Ocean. Initially, the energy of such a wave is distributed throughout the water column, and surface perturbation is small. Only when the water grows shallow, near the coast, does the wave emerge on the surface as a tsunami — the name is Japanese for 'harbour wave'. In this case, the wave hit Indonesia and Thailand within an hour, and then Sri Lanka and India, ultimately reaching as far as eastern Africa.

Michael Hopkin



The tsunami driven by an oceanic earthquake caused widespread destruction, as shown by these views of Banda Aceh, Indonesia, before and after the disaster.

always on the agenda," says Vasily Titov, a tsunami researcher at the Pacific Marine Environmental Laboratory in Seattle, Washington. But he says that it has been difficult to raise the money for a monitoring system. "Only two weeks ago it would have sounded crazy," he says. "But it sounds very reasonable now. The millions of dollars needed would have saved thousands and thousands of lives."

The most recent comparable event in the region took place in 1883 (see 'Tsunamis: a long-term threat', right). In contrast, earthquakes in Chile in 1960 and Alaska in 1964 led to the creation of a reasonably sophisticated tsunami warning system in the Pacific Ocean. Two international tsunami warning bodies exist under UNESCO's Intergovernmental Oceanographic Commission (IOC): the International Coordination Group for the Tsunami Warning System in the Pacific, known as ITSU, and the International Tsunami Information Center based in Hawaii. They get by on annual budgets from the IOC of about US\$40,000 and \$80,000, respectively, which are supplemented by grants from nations on the Pacific rim.

Displacement data

To predict a tsunami with any useful time advantage, researchers say, data on small changes in sea level and pressure have to be collected directly from the floor and surface of the ocean. The strength of the event depends on the displacement of the ocean floor, not on the strength of the earthquake.

Some buoys that could provide such data are already in place in the Indian Ocean. And only a few weeks before the tsunami struck, members of ITSU were talking about how these could be adapted for use in a tsunami-

warning system, says Peter Pissierssens, head of ocean services at the IOC.

Within 20 minutes of the earthquake, at least three monitoring stations in the United States had detected it, initially estimating its magnitude to be around 8. The United States Geological Survey (USGS) circulated the information to about 100 people, mostly its own researchers and senior officials, within 16 minutes, and sent a more detailed bulletin to a list of external contacts, including the US Department of State, after an hour. The USGS has no responsibility for tsunami

monitoring and its statement did not mention the risk of such an event.

The Hawaii-based Pacific Tsunami Warning Center (PTWC), meanwhile, sent out a bulletin to its regular circulation list, noting that the event presented no tsunami risk in the Pacific. According to Laura Kong, director of the International Tsunami Information Center, "let's keep an eye on it" was the prevalent attitude that night. "At that point, none of us expected anything like what we have seen," says Charles McCreery, director of the PTWC and deputy chair of

Tsunamis: a long-term threat

Last month's tsunami tragedy, shocking as it was, had ample historical precedent. On 1 November 1755, for example, a fire following an earthquake destroyed two-thirds of Lisbon, Portugal. In panic, the population sought shelter near the shoreline, only to be hit by waves said to be as high as houses. More than 60,000 people died.

Devastating tsunamis are known in historical times to have affected the populated coasts of Papua New Guinea, Japan, Hawaii, Crete, Sicily and the Crimea — to name just a few. In the Pacific region, where 80% of all tsunamis occur, a 1947 analysis indicated that seismic sea waves higher than 7.5 metres occur on average every 15 years¹. Records going back to 684 BC refer to four Pacific tsunamis higher than 30 metres.

Outside the Pacific, tsunami frequencies have been studied in some detail only for the Aegean and Black Sea regions. Records there reveal that the coastal and surrounding areas of Turkey have been affected by more than 90 tsunamis over the past 3,000 years².

For most other areas, information concerning

the return periods of tsunamis is scarce. A rough comparison of tsunami frequencies in different parts of the globe was done in 2000 by the London-based Benfield Hazard Research Centre, as part of its Tsunami Risks Project. The resulting risk analysis estimates the return periods of 10-metre waves to be about 1,000 years for the North Atlantic and Indian oceans, southern Japan and the Caribbean, 500 years for the Philippines and the Mediterranean Sea, 250 years for Alaska, South America and Kamchatka in eastern Siberia, and less than 200 years for Hawaii and the southwest Pacific.

The south Asian disaster will have a "huge effect" on instigating more thorough risk assessments, predicts Bill McGuire, a volcanologist and director of the London research centre, as well as encouraging preventive measures in threatened regions.

Quirin Schiermeier

♦ www.nerc-bas.ac.uk/tsunami-risks

1. Heck, N. H. *Bull. Seismol. Soc. Am.* **37**, 269–286 (1947).
2. Altinok, Y. & Ersoy, S. *Nat. Hazards* **21**, 185–205 (2000).

India pledges to fund alert system in wake of disaster

India's government and scientific establishment have been heavily criticized for failing to provide warning of a tsunami that drowned at least 12,000 people on the nation's eastern coast.

Newspapers and opposition spokesmen have asked why a country with India's scientific resources couldn't better prepare for such an event. Ministers immediately pledged up to US\$29 million to build a tsunami-monitoring system, and promised to seek more cooperation with the Pacific Tsunami Warning Center in Hawaii.

"This is not a knee-jerk reaction. We are very serious," science and technology secretary Valangiman Ramamurthi told *Nature*. "We are going to have a brain-storming meeting this month to decide how we should proceed and we have invited experts from the United States," he said. In response to criticism, he added: "We cannot join a Pacific network as India is not in that region. And you do not make heavy investment to warn against something that happens once in a century."

The ocean development secretary, Harsh Gupta, told a press conference in New Delhi



Relief centres in India have been inundated with people in need of food and aid.

that there was no record of a tsunami ever hitting the Indian coastline, even as other government ministers acknowledged such events in 1833 and 1883.

"No government thought of it," says science minister Kapil Sibal. "The last recorded tsunami was in 1883. It was not in the horizon of our thoughts." India now plans to install a network of 10 to 12 seafloor pressure sensors to be imported from the United States, as well as several floating

sensors on ocean buoys, linked to an Indian geostationary satellite.

Critics say that the tragedy exposed a major weakness in the current system, which authorizes only the Indian Meteorological Department to put out hazard alerts. "Data were pouring into our lab but we cannot issue alerts even if we can analyse the data for tsunami potential," says one researcher at the National Geophysical Research Institute in Hyderabad.

They also want to know why the Indian air force, whose base in Car Nicobar Island was submerged by tides an hour before the waves hit the mainland, failed to provide any public warning.

The tsunami spared India's main rocket launch site at Sriharikota Island, 80 kilometres north of Chennai. But it damaged cooling water pumps at a nuclear power station at Kalpakkam, leaving staff with very little time to shut down the plant safely. "The tsunami factor was not taken into account," says Anil Kakodkar, chairman of the Atomic Energy Commission. "From now on, it will be factored in."

K. S. Jayaraman, New Delhi

ITSU. "We expected a local tsunami at most."

At 2:04 GMT, the PTWC put out another bulletin revising the quake up to magnitude 8.5. Because there was no information about sea levels in the area, the existence of a tsunami was merely hypothetical, but staff were worried enough to begin looking for numbers to call in Asia.

Communication breakdown

According to Kong, the team tried and failed to reach colleagues in Indonesia. Australia was contacted, although to little avail, as that country experienced only half-metre waves. It was not until 3:30 that the team in Hawaii saw news reports on the Internet of casualties in Sri Lanka. The wave had already crossed the ocean, to devastating effect.

Kong says that without a predetermined communication plan, warning efforts were doomed from the start. But she adds that the PTWC will in future directly contact the US state department, which can communicate risks to any nation, at any time.

Indonesian seismologists initially underestimated the strength of the earthquake, according to local news reports. And although officials there had very little time in which to act, an instrument that could have helped warn them of the approaching wave was transmitting its information to a dead phone line, according to a senior Indonesian seismologist (see *news@nature.com* doi:10.1038/news041229-4; 2004).

Efforts over the years to get an Indian Ocean warning system in place have made little progress in the face of national governments' reluctance to invest in them. In 2003, a working group on the Tsunami Warning System in the Southwest Pacific and Indian

Ocean was established within ITSU. But Pissierssens says that the first chair of the group, a representative from Indonesia, left soon after his appointment and that the group then split into two according to region.

Phil Cummins a seismologist at Geoscience Australia in Canberra agreed to write a position paper for the group on tsunami risk in the Indian Ocean. "I am still in the process of writing that paper," he says. "No one else was 100% convinced that we should worry and that included me, I've got to admit."

According to Pissierssens, UNESCO will now make an observation system in the Indian Ocean a priority. "The first thing we will do is send out a survey team in January or February," he says, "and then we want to set up

a conference in the area." Needless to say, there is little reluctance now to accept the need for the system. The UN International Strategy for Disaster Reduction has also said that one should be built within a year. And the Indian government, under intense domestic pressure for its failure to warn people on its eastern coast, said it would spend up to US\$29 million to build a system itself (see 'India pledges to fund alert system in wake of disaster', above).

Nicole Rencoret, spokeswoman for the UN's disaster-reduction branch, notes that early warning systems could watch for other natural disaster risks, as well as tsunamis. "There has been an enormous amount of focus on tsunamis, but we need to take a multihazard approach," she says.



Turning tide: the waters of the Indian Ocean tsunami recede after battering the coast of Sri Lanka.

Eli Lilly posts clinical trials data online for top-selling drugs

Washington Data from clinical trials of blockbuster drugs have been made public by pharmaceutical company Eli Lilly.

The Indianapolis-based firm launched the online Lilly Clinical Trial Registry (www.lillytrials.com), which contains trial data on the company's products, in December. So far, the website features information on eight drugs, including the antidepressant Prozac and Zyprexa, a multibillion-dollar drug for schizophrenia.

Lilly says that it aims to expand the site to include clinical trials data on all of its products by the middle of the year.

Some critics say that the move is an attempt by the drug industry to pre-empt congressional legislation proposed last October, which would oblige companies to make all trials data public — including those for drugs that didn't work out.

"Having a few companies volunteer to post clinical trials data is not a substitute for general disclosure by everyone," says Sidney Wolfe, director of the health research group at Public Citizen, a health-advocacy group based in Washington DC. "The remedy is legislation requiring

companies to set up clinical trial registries of every trial they have done and to fully disclose results within a fairly short period of time."

California gets set to expand stem-cell research

San Francisco California's new stem-cell institute got off the ground last month. The oversight panel for the California Institute for Regenerative Medicine, which will

distribute \$3 billion for human stem-cell research within the state over ten years, met for the first time on 17 December at the University of California, San Francisco.

Robert Klein, a real-estate developer and the prime mover behind the institute (see *Nature* 432, 135; 2004), was unanimously elected chairman. Vice-chair is Edward Penhoet, founder of the Emeryville biotechnology company Chiron and president of the Gordon and Betty Moore Foundation. Other panel members include

Antimatter-spotting balloon takes pole position

Washington A polar balloon has completed an eight-day search for particles of antimatter. The study could provide evidence for Hawking radiation — an elusive stream of particles emitted by black holes.

The high-altitude experiment, BESS-Polar, which returned to ground in Antarctica on 21 December, was searching for antiprotons. Cosmologist Stephen Hawking of the University of Cambridge, UK, has predicted that black holes should emit these particles with a specific range of energies. Researchers at Japan's High Energy Accelerator Research Organization (KEK) in Tsukuba, together with colleagues from elsewhere in Japan and the United States, are now analysing data from the flight.



"Earlier, shorter flights provided hints of the signature of Hawking radiation," says Akira Yamamoto, a physicist at KEK. "With a longer flight and a great harvest of antiprotons we might be able to show that Professor Hawking is right."

BESS/GSFC/NASA

university, biotechnology and patients' advocates. Plans to discuss royalties from future inventions were postponed after complaints of inadequate public notice. "If there is ever any question, we will err on the side of caution," Klein said.

Huygens goes it alone for last lap to Titan

London The Huygens probe is now on a collision course with Titan, Saturn's largest moon, having separated from the Cassini spacecraft at 02:00 GMT on Saturday 25 December.

The European Space Agency's probe will dive through Titan's atmosphere, measuring its gases and weather. If Huygens survives the landing, its instruments may also send back details of Titan's surface during the two hours it is in contact with Cassini.

The probe will remain dormant until four hours before it arrives at the edge of Titan's atmosphere on 14 January 2005, when three electronic alarm clocks will rouse it from slumber and release its parachutes. The team hopes that Huygens will splash down safely into liquid ethane, where it will be able to use its sonar to plumb the depths of the alien sea. However, it might crash land in a pool of tarry hydrocarbons or even on icy rock.



Forbidden fruit: cannabis production for research trials has been banned by drugs agency.

Grow-your-own cannabis gets the thumbs down

San Diego A US researcher has at last received a reply to an application to grow marijuana for pharmaceutical research. After a wait of more than three years the answer is "no".

Lyle Craker, a botanist at the University of Massachusetts at Amherst, applied to the US Drug Enforcement Administration (DEA) for a permit to grow his own plants to obtain high-quality marijuana for private trials funded by the Multidisciplinary Association for Psychedelic Studies (MAPS) in Florida. Last July, frustrated by the long delay, MAPS

sued the DEA and other federal bodies for allegedly stalling the permit request (see *Nature* 430, 492; 2004). The request was finally turned down on 10 December.

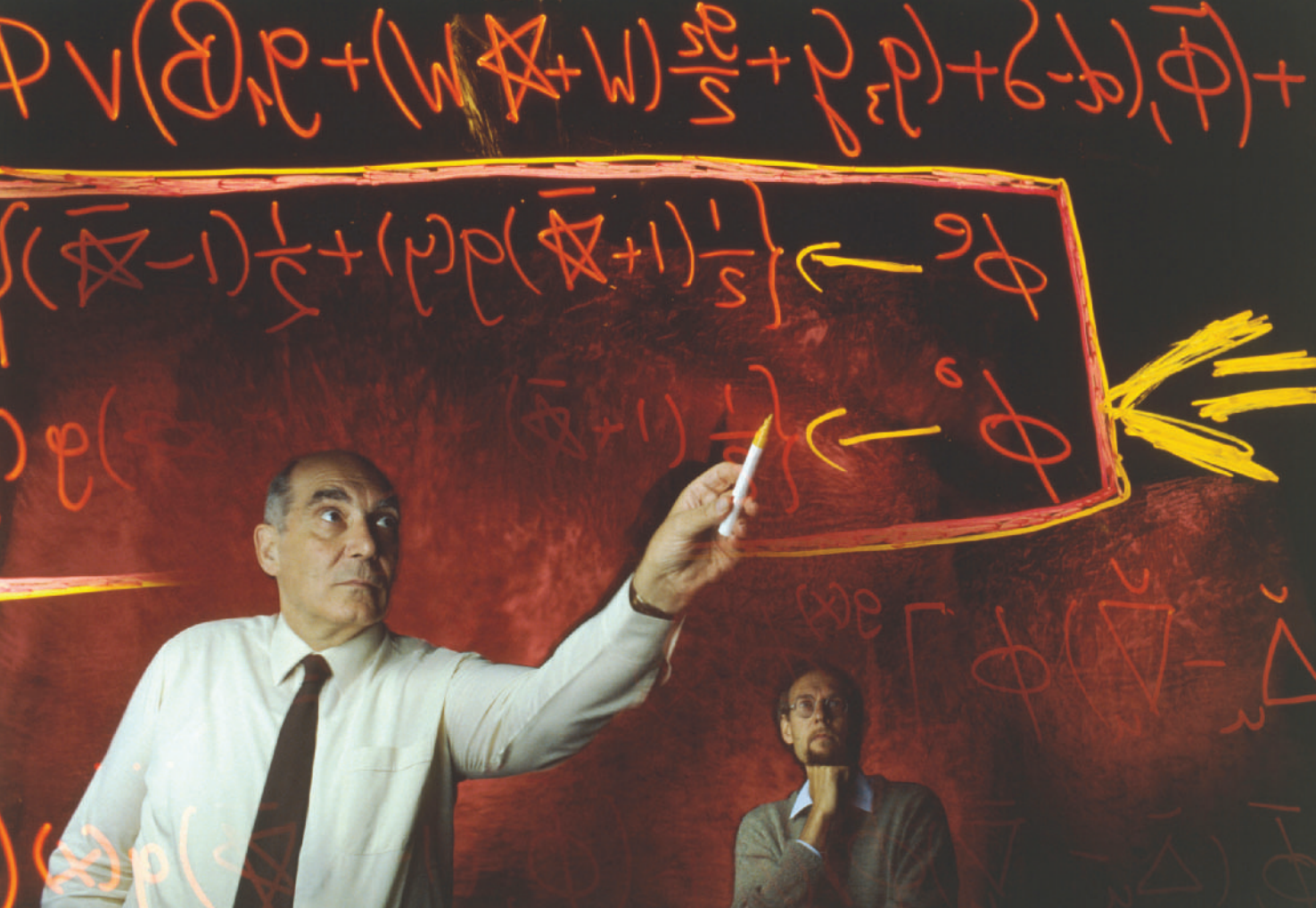
MAPS president Rick Doblin says the organization will appeal against the denial.

No smoking gun to link firearms and crime rate

Washington Despite widespread gun violence in the United States — 10,000 homicides a year, 17,000 suicides and 75,000 injuries — solid research on measures to prevent such incidents is sorely lacking, says a new report from the National Academy of Sciences.

Firearms research has engendered heated debate, with researchers such as John Lott of the American Enterprise Institute claiming that gun ownership helps deter crime. But the academy panel concluded there is "no credible evidence" to support this.

The one dissenting voice to this conclusion was that of panel member and conservative scholar James Wilson, emeritus professor of public policy at the University of California, Los Angeles. "I think Lott's results are significant," he said. Charles Wellford, the University of Maryland criminologist who chaired the panel, emphasized "data limitations" as the main problem to be addressed.



So, what's your theory?

One hundred years ago, when Albert Einstein penned his era-defining papers on brownian motion, the photoelectric effect and special relativity, he was just 26 years old. Reaching scientific greatness at such a young age was exceptional then and may be even harder today. But when looking at physics in the twenty-first century, there's still much to discover by asking a young physicist: what's your theory?

In the following pages, *Nature* offers a glimpse into the lives of four young theorists (all under 35) who are making waves in their chosen fields.

In Einstein's youth, the focus of theoretical physics was in Europe: Niels Bohr was in Copenhagen, Max Planck was in Berlin. Today it is harder to find the centre of the theoretical universe — collaborative research is increasingly international, and most theorists, who need little more than a laptop, can work anywhere.

But it seems that many young theorists opt to spend their formative years in the United States. Although US enrolment of foreign graduate students has fallen in recent years, they still make up about half

of the total in physics, and of these some 40% are theorists. Not all of these students will stay in the United States — many, including three featured here, will head for good positions back home.

All of our interviewees share a willingness to push big ideas forwards while also asking how — and how soon — they can be tested. Einstein had to wait just a few years for predictions from his 1915 general theory of relativity to be confirmed by observations of a solar eclipse. But many theorists finish their careers without seeing any experimental check on their ideas. Our young theorists not only know that they must think big, but also that they must pit their wildest theories against reality.

Over the next few years, many theorists will be directing their attention to the world's largest particle accelerator, currently being built near Geneva in Switzerland, for experimental confirmation of their ideas. Here, physicists hope to find some support for exotic notions

such as extra dimensions (see 'In search of hidden dimensions', page 10). Elsewhere, missions such as the Planck satellite will provide data on the early Universe that may help to shore up theories about the moment of creation (see 'The long-distance thinker', page 12). The hoped-for construction of the first quantum computer would test fundamental ideas about the quantum world (see 'A theorist of errors', page 9),

whereas the creation of exotic materials is continually pushing our knowledge of electron behaviour (see 'Can electrons do the splits?', page 11).

As we look back at the past 100 years and celebrate the World Year of Physics in 2005, there are still plenty of twentieth-century puzzles that theorists would like to complete, but there is also the tantalizing possibility that something totally unknown is just around the corner. It will

be young theorists such as those featured here who will tackle these challenges. ■

Sarah Tomlin edits News Features for *Nature* from New York.



Young star: Einstein was only 26 when he published three seminal papers.

A theorist of errors

Growing up on Einstein Street in Haifa, Israel, Dorit Aharonov was perhaps destined to study physics. But she pursued other interests before finally settling on quantum computation. Haim Watzman reports.

To enter Dorit Aharonov's office is to experience a sudden transition between order and disorder. The corridors of the computer-science building at the Hebrew University of Jerusalem are stark, white and neat. Aharonov's office is a jumble of red-and-orange patterned cushions, article reprints and wicker furniture. It's an appropriate setting for a theorist who has proved that when disorder reaches a certain level, the physics of the quantum realm switches into the classical domain of the world we see every day.

Aharonov devotes herself to the theory behind quantum computers. As-yet unbuilt, these machines would harness the power of quantum mechanics to perform tasks that defeat conventional computers — such as factoring large numbers. Aharonov, now 34, has already made important contributions to this goal by showing that a quantum computer could perform reliably and accurately despite a 'noisy' environment.

Physics runs strong in Aharonov's family. Her uncle, Yakir Aharonov, is a physicist at Tel Aviv University, and her father is a mathematician who taught her the beauty of numbers when she was little. She later chose physics and mathematics for her undergraduate studies, but the quantum world did not initially capture her imagination. She wanted instead to use physics to study the brain.

A chance encounter

"I wanted to solve the problem of consciousness," she recalls. But she began to think that the problem was still beyond the reach of today's science. "Then, one day, at a wedding, a friend asked me for advice about what direction to take in the study of the brain. I advised him to check out what people in computer science were doing," she says.

Realizing she should take her own advice, Aharonov went to the Hebrew University's computer-science building to find someone to talk to. She was directed to Michael Ben-Or and, as she knocked on his door, she says that she had a strong feeling something important was going to happen. It did. Ben-Or told her about quantum computation. "It fascinated me. It was mathematics, physics and philosophy all in one package," she says.

Back then, in 1994, the problem facing theorists such as Ben-Or was how to prevent

a quantum computer from crashing. All computers make errors when they operate, but quantum computers are more susceptible to failure. This is because the quantum states on which calculations depend are very delicate: complex phenomena, such as the spin states of atomic nuclei, can store quantum information but this data can easily be lost if the particles interact with their surroundings. A computer can never be perfectly isolated from its environment, so there will always be 'noise' in the system and, inevitably, errors will arise. Moreover, correcting such errors is almost as difficult as doing the calculation in the first place. So will it ever be possible to do a reliable quantum calculation?

"That was the problem I posed to Dorit," says Ben-Or, who became Aharonov's dissertation supervisor and later her collaborator. Working with Ben-Or, Aharonov proved that at a constant but low level of system noise, a quantum computer can still produce accurate results¹.

"I consider her to be one of the most outstanding young people in this field," says Peter Zoller, a theoretical physicist at the University of Innsbruck, Austria. Zoller wants to build a quantum computer, and he says that Aharonov has been instrumental in laying the theoretical foundations on which a real machine could be constructed. As well as her work on error tolerance, he cites an important proof² Aharonov developed with Oded Regev and others while working at the University of California, Berkeley. The proof showed that two existing models for quantum computing are actually equivalent and, as a result, made writing quantum algorithms easier.

While at Berkeley, Aharonov extended her work on computers to address a fundamental puzzle presented by quantum mechanics — why its laws are evident in the world of elementary particles, but not in everyday life. At what point does the world switch from looking quantum to looking

classical? Is it simply a matter of scale?

Aharonov showed that for many noisy quantum systems, there is a level of noise above which a transition to classical behaviour is inevitable. Such transitions are much sharper than expected from other theories that predict a gradual shift away from quantum behaviour³.

Ben-Or says that what sets Aharonov apart is her boldness. As a graduate student she was not shy about contacting leading figures in the field to discuss their work, he recalls. Zeph Landau, a mathematician at the City College of New York who collaborated with Aharonov on the model equivalence paper, says that she is focused but not single-minded, finding time to discuss other pursuits.

Aharonov says that balancing life and work is essential to her research. Like many theorists, she says that she has her best ideas when not thinking about work at all. Her daily yoga session is particularly rewarding, she says: "It disperses the fog. My intuition becomes sharper. When there is less struggle, ideas become clear."

Eastern ideas about the interconnectedness of everything also influence her work. For instance, Aharonov is not fixated on the actual construction of a quantum computer. "The most interesting thing that might come out of an attempt to build one is the discovery that we can't do it," she says. By failing, she adds, we might discover some entirely new physics. ■

Haim Watzman is a freelancer based in Jerusalem, Israel.

1. Aharonov, D. & Ben-Or, M. Preprint at <http://xxx.lanl.gov/quant-ph/9611025> (1996).
2. Aharonov, D. et al. Preprint at <http://xxx.lanl.gov/quant-ph/0405098>, (2004).
3. Aharonov, D. *Phys. Rev. A* **62**, 062311 (2000).



Relaxed: Dorit Aharonov finds her yoga inspires her theories.

In search of hidden dimensions

So far, string theory has defied experiments, but Nima Arkani-Hamed thinks he has found a way to put the idea to the test. Geoff Brumfiel finds out how.

Ask most theorists when they think their calculations will be tested experimentally and you'll be told "decades" or sometimes, more honestly, "never".

But ask Nima Arkani-Hamed, a physicist at Harvard University, and he will give you a far closer date: 2008. That is when the first results from the Large Hadron Collider, the world's most powerful particle accelerator, are expected to be released by CERN, the European particle-physics laboratory near Geneva, Switzerland. And if Arkani-Hamed's predictions are correct, then that is when an experiment will detect the first evidence to support string theory — a vision of the cosmos that has never been verified experimentally. "The field is going to turn on what happens at the collider," he says.

Pacing his sparse Harvard office, the 32-year-old physicist drinks no less than six cups of espresso during our hour-and-a-half interview, as he tries to explain why he thinks string theory can now be tested.

String theory emerged in the 1980s as a way to answer questions that still baffle modern physics, such as why is gravity so much weaker than other fundamental forces? By imagining that everything is composed entirely of strings ten billion billion times smaller than atomic nuclei, theoretical physicists were able to create a model of the Universe that unified all fundamental forces into one, and described most of the particles we see today. Unfortunately, these strings are far too small to be detected by even the most powerful particle accelerators. And so, critics say, they are more philosophy than physics.

Arkani-Hamed's ideas have very little to do with strings themselves. Instead, he is hoping to detect the extra dimensions predicted by the theory, which, like the strings, are thought to be vanishingly small. But in 1998, Arkani-Hamed and his colleagues published calculations showing that some of these extra dimensions might be

"String theory just seemed like abstruse junk to me. What I really liked was physics that explained things about the world around me."

— Nima Arkani-Hamed



String fellow: Nima Arkani-Hamed hopes that particle-collision experiments will show that gravity leaks into other dimensions.

as large as a millimetre (N. Arkani-Hamed, S. Dimopoulos and G. Dvali *Phys. Lett. B* **429**, 263–272; 1998). Such large dimensions, they argued, have escaped detection because everything we know — except for gravity — is confined to the three dimensions of space and one of time. But gravity, they think, might be able to seep into these extra dimensions. This would explain why it seems so weak to us. And, as a result, unexpected variations in gravity could allow researchers to detect the hidden dimensions.

Leaking away

"It was a watershed event in the field," recalls Joe Lykken, a theoretical physicist at Fermilab near Chicago in Illinois. Suddenly, a theory that most thought could never be tested was within experimental reach. Some groups rushed to look for deviations in gravity at small scales. So far, they have nothing to report, but the hope created by Arkani-Hamed's work is enough to win him wide praise. "The word 'genius' is overused, but I think it is easily applicable in the case of Nima," says Savas Dimopoulos, a Stanford theorist and one of Arkani-Hamed's collaborators.

The son of two Iranian physicists, Arkani-Hamed was born in Houston, Texas, and grew up in Boston. After the Iranian revolution of 1979, his family returned to their homeland, but as religious fundamentalists took over the government, his father was forced to go underground and the family eventually had to flee across the border to

Turkey. By 1982, Nima was living in Toronto, Canada.

Recalling his early life, Arkani-Hamed says that his time in Iran was largely a positive experience. "The strange thing is that I have mostly wonderful memories," he says. If anything, he adds, it taught him to worry less about what others thought of him. "Given that so many aspects of my life have been unusual, I've never had a problem with feeling different or being different or doing different things."

As a child, Arkani-Hamed loved physics, but he initially disliked almost everything about string theory. "String theory just seemed like abstruse junk to me," he says. "What I really liked was physics that explained things about the world around me."

That changed when he began studying quantum field theory at the University of

Toronto. At first, this complex theory — which underlies high-energy physics and much of string theory — seemed too arcane, but as he studied it more carefully, he found a level of order and explanation far beyond anything he had learned before. "Clearly, there was something very deep going on," he says.

It captivated him, and by the time he finished graduate school in 1997, he knew he wanted to try to make string theory experimentally verifiable. He found an ally and mentor in Dimopoulos, who has devoted his career to seeking testable versions of string theory. "We believe that the only way to make progress is to take an idea, and push its consequences to find observations," Dimopoulos says.

These days, in late-night phone calls and frequent e-mails, the two are thinking about what might emerge at the Large Hadron Collider. Their current calculations show that some of the energy created by particle collisions in the machine could escape into extra dimensions, carried off by leaking gravity, if those dimensions are large enough. The result would be an apparent violation of the conservation of energy — a dramatic sign that string theorists are on the right track.

Then again, they might not be. "You can spend ten years of your life and every idea you come up with can be wrong, and that's gratifying in its own way," Arkani-Hamed says. But, he adds, as he reaches his caffeine-fuelled conclusion: "If this thing turned out to be true, it could be the biggest discovery in science in, say, 300 years."

Geoff Brumfiel is *Nature's* Washington physical sciences correspondent.

Can electrons do the splits?

The electronic behaviour of some forms of matter doesn't match theory. Geoff Brumfiel meets Senthil Todadri, a man who wants to change our view of how electrons behave.

Right from the start, it was clear that Senthil Todadri was no ordinary graduate student, says Subir Sachdev, a professor of physics and Senthil's adviser at Yale University. On his first day, Senthil made several observations that forced Sachdev to rethink his work. "He himself didn't understand the depth to which he understood things," Sachdev says. That first day's work was enough to win Senthil co-authorship on the group's next paper.

Senthil (who grew up with no last name, but adopted his father's name, Todadri, when he came to the United States) is the son of a banker in the Indian city of Chennai. "I was going to work in a bank just like my father, but got more and more sucked into maths and science," he says. "My family considered it a bit bizarre when I decided to take up physics." After completing his undergraduate degree at the Indian Institute of Technology in Kanpur, he began his graduate studies at Yale in 1992.

Since his first day on the job, Senthil, now 34, has continued to make waves in condensed-matter physics, a field whose theoretical underpinnings are in upheaval. Since the early 1980s, experimentalists have uncovered dozens of materials that defy the present theory of how electrons behave in solids — often referred to as the Fermi liquid model. Senthil is helping to build a new theoretical framework that could explain these exotic materials, the most alluring of which are superconductors (they have no electrical resistance) at temperatures that exceed those predicted by current models.

Deep divisions

One of the more unconventional ideas Senthil has pursued is that an electron added to a material can 'split' under the right circumstances, so that a fraction of its charge goes one way, and a bit of its 'spin' the other. "It's a pretty dramatic thing if you think about it because an electron is supposed to be a fundamental particle," Senthil says. The electron loses its identity in the collective behaviour of other electrons in the solid, he explains. As the electron's fundamental characteristics of charge and spin are shared among the other electrons, it essentially splits into fractional particles of spin and charge.

The idea of electron splitting has been around since the 1980s, but in the context of high-temperature superconductors it has been taken seriously only in the past five years or so — thanks in part to work by Senthil and



In a spin: Senthil Todadri is interested in whether electrons can split themselves up inside solids.

his colleague Matthew Fisher of the University of California, Santa Barbara. In 2001, Senthil and Fisher proposed a novel way that their ideas could be tested against the behaviour of certain high-temperature superconductors¹.

Not long after, an experimental group at Stanford set out to search for the fractional charges². Unfortunately, the team failed to find the exotic behaviour predicted by Senthil and Fisher, says Piers Coleman, a theoretical physicist at Rutgers University in New Jersey. "It was a nice idea that didn't work out, but that's okay," says Coleman. "Good science has interesting proposals that can be tested. I think everyone regarded the work they did as extremely interesting and very stimulating."

Sachdev, who has been a co-author on Senthil's work, agrees. "Since his paper appeared a few years ago, we've found that those ideas have turned out to be remarkably powerful." Although the version of the theory proposed by Senthil and Fisher was proved wrong for high-temperature superconductors, they continue to explore ways in which the collective properties of a solid's electrons can shape its behaviour.

Most recently, Senthil and his collaborators have made impressive progress in describing quantum phase transitions — sudden shifts in a material's behaviour that are caused by the quantum fluctuations of its electrons at a temperature of absolute zero³. Once again, this theory depends on

fractional parts of electrons appearing briefly at the point where the material changes from one state to another.

Homeward bound

This January, Senthil will leave his position at the Massachusetts Institute of Technology and return to India, where he will lead a theoretical group at the Indian Institute of Science in Bangalore. The reasons for the move are personal and professional, he says. It will allow him, his wife and his young daughter to be closer to the rest of their family.

"I don't know what it is really going to be like," he says. But the quality of Indian physics has been steadily improving over the past few decades, he adds. One outstanding question is whether Senthil will be able to recruit the high-quality graduate students that form the backbone of any good theory group. "I'm hoping it will be possible to get good students and postdocs in India, but I don't have firsthand experience," he says.

Despite some setbacks and uncertainties, Senthil remains confident that within a decade an entirely new set of theories will be developed that can explain even the most bizarre of materials. "We're just starting to glimpse an entirely new world inside solids," he says. "It's a great time for condensed-matter physics because a lot of the stuff we teach in textbooks needs to be revisited."

Geoff Brumfiel is *Nature's* Washington physical sciences correspondent.

1. Senthil, T. & Fisher, M. P. A. *Phys. Rev. Lett.* **86**, 292–295 (2001).
2. Bonn, D. A. et al. *Nature* **414**, 887–889 (2001).
3. Senthil, T., Vishwanath, A., Balents, L., Sachdev, S. & Fisher, M. P. A. *Science* **303**, 1490–1494 (2004).

The long-distance thinker

Martin Bojowald is on a journey back in time to see what happened during the Big Bang. Quirin Schiermeier tags along for the ride.

The journey southwest from Berlin to Golm, a small village near Potsdam, is a 90-minute train trip to the end of the world. Or that is how it seemed on a misty December morning. Outside Potsdam the only view from the window is farmland stretching to the horizon, until an ultra-modern glass building looms out of the fog.

This think-tank in the middle of nowhere is the Max Planck Institute for Gravitational Physics, often called the Albert Einstein Institute. As might be expected, it is home to theorists who are struggling with physics' deepest questions. How did the Universe begin? What will be its fate? And what happens to time, space and matter at these extremes?

The forlorn landscape outside rather suits Martin Bojowald, a 31-year-old German theorist, who admits he spends most of his time staring into space. Except when writing papers or e-mails, he hardly uses a computer — and he does most of his deep thinking at home, where he feels less self-conscious about his apparent lack of activity.

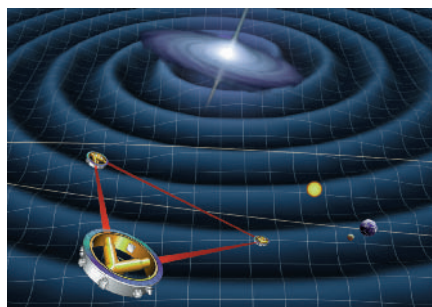
Bojowald is a disciple of loop quantum gravity, a theory of gravity at the smallest of scales, which physicists can use to look inside black holes or back to the first few moments of the Universe.

Loop quantum gravity is a way to reconcile general relativity — Einstein's theory of how gravity shapes the cosmos — with our quantum picture of the atomic world. Gravity, alone among the four fundamental forces of nature, seems not to respect the rules of quantum physics. Loop quantum gravity tries to address this directly, by rewriting Einstein's equations within a quantum framework. A popular alternative route to 'quantum gravity' is provided by string theory, which has its roots in particle physics, and postulates that everything in the Universe is made of unobservable vibrating strings.

Any decent theory that claims to unite general relativity with quantum theory should be able to fix some of the cosmological puzzles unsolved by general relativity.



Martin Bojowald, whose idol is Einstein, hopes that his ideas on the Big Bang will gain some support from NASA's LISA mission (below).



One enduring mystery is figuring out what happened during the Big Bang — the cosmic event that about 15 billion years ago gave birth to a hot, dense fireball and eventually, stars, galaxies and humans. Although Einstein's equations can describe much of the Universe's history, they break down the closer we get to this moment of creation.

Off with a bang

Conventional wisdom says that the Big Bang was the start of everything, including time, so questions about the Big Bang itself, or what came before, don't make sense. Or so we're told. But the breakdown in the laws of physics — the singularity problem — limits what we know about the starting conditions of the Universe. So it leads to arbitrary assumptions, such as an early period of rapid expansion (inflation), to get the Universe to where it is now.

It is in part thanks to Bojowald that a cosmology based on loop quantum gravity has become a respected, albeit controversial, notion. "Martin has opened the door to the possibility of calculating the predictions of loop theory for cosmology, and determining whether they can be tested against observa-

tions," says Roy Maartens, a cosmologist at the University of Portsmouth, UK.

In the loop quantum universe everything is quantized, or discrete, including time. Space can be chopped up into discrete 'cubes', just 10^{-99} cm^3 . One cube would equal the smallest unit of space, but it is not 'empty' space; each cube incorporates space, time and matter in the form of intersecting 'loops'.

"This has few consequences for our understanding of the real world," says Bojowald. These loops operate on scales far outside our experience. "But the discreteness of loop theory makes it much easier mathematically and conceptually to come to terms with the early Universe," he says.

In the loop

Although the loop language is complex, the maths behind the theory is elegant. Bojowald has created a framework in which physical laws do not break down at the Big Bang singularity (M. Bojowald *Phys. Rev. Lett.* **86**, 5227–5230; 2001). His results suggest that at extremely small scales, quantum gravitation can be repulsive, which prevents the collapse of space-time into a singularity. This effect, which would contradict general relativity, might be a consequence of the quantization of Einstein's equations, Bojowald says.

Freed from the singularity, Bojowald can now look back to a time 'before' the Big Bang. He finds an inverted universe on the other side — a mirror-image of ours — expanding outwards as time runs backwards.

Bojowald's model also provides tantalizing insight into how inflation occurs (M. Bojowald *Phys. Rev. Lett.* **89**, 261301; 2002). A gravitational repulsion not only prevents the collapse of a contracting universe, he believes, but also pulls apart an expanding one. Maartens cautions that this idea has some way to go before it is fully convincing. But that long road doesn't intimidate Bojowald, who is a long-distance runner both in real life and in science.

"In the beginning, there was a lot of criticism," Bojowald says. "But things have changed, and meanwhile many cosmologists have got very interested in loop equations."

Bojowald hopes that data from the European Space Agency's 2007 Planck mission will provide indirect backing for his ideas. This satellite will test theories of the early Universe by looking at the radiation left over from the Big Bang. After 2011, data from NASA's Laser Interferometer Space Antenna could reveal a quantum gravity effect from the early Universe in its observations of ripples in space-time.

In the meantime, says Sean Carroll, a theoretical cosmologist at the University of Chicago, Illinois, string theory remains the more popular theory, given that it has solved many problems related to quantum gravity. "But," he adds, "any alternative concept is welcome and needs to be taken seriously." ■

Quirin Schiermeier is *Nature's* German correspondent.

Destructive fires are not just Indonesia's problem

Logging, urban expansion and lawlessness fan the flames throughout Borneo.

Sir—Your News Feature “Borneo is burning” (*Nature* 432, 144–146; 2004) links the mismanagement of peat swamp forests in Central Kalimantan to the appallingly destructive fires that leave the region blanketed in haze and release massive amounts of carbon dioxide whenever there is a substantial drought.

This informative account of the environmental problems associated with Suharto's Mega Rice Project, and recent attempts to rectify them, unfortunately reinforces a misperception that these fires are largely an Indonesian problem, and thus that their ultimate causes lie in the particulars of Indonesian politics.

During the 1997–1998 El Niño event described in the News Feature, drought and fires were widespread in the Malaysian states of Sarawak and Sabah and the

independent state of Brunei Darussalam, as well as in Kalimantan. Also, fires were not restricted to peat swamps, as the Indonesian experience suggests, but occurred in agricultural areas, logged forest and even primary rainforest — although peat fires tend to burn for longer and release much larger amounts of smoke and carbon dioxide.

As someone who witnessed these fires and experienced the debilitating effects of the resulting haze, the focus on Indonesia has always disappointed me. If the problem of burning in Borneo and elsewhere in southeast Asia is to be properly addressed, governments and donors in the region must first recognize the widespread nature of the problem.

Indeed, the principal causes are not difficult to identify: they are environmental

mismanagement, in particular the development of peat-swamp areas for agriculture (as mentioned in the News Feature), oil-palm plantations or urban expansion; increased access to formerly remote areas, often as a result of logging; and lack of law enforcement because of governments' reluctance or inability to assert authority at a local level.

Until these problems are addressed, fire and haze will continue to plague the region whenever there is a prolonged drought. The ongoing massive destruction of natural environments and associated carbon dioxide emissions make this a global issue of considerable urgency.

Rhett D. Harrison

*Smithsonian Tropical Research Institute,
Apartado Postal 2072, Balboa Avenue,
Ancon, Republic of Panama*

Fighting future fires with fairness

Sir—In your Editorial “Burning issues” (*Nature* 432, 131; 2004) you argue that international climate treaties, such as the Kyoto Protocol, should provide incentives to reduce greenhouse-gas emissions from peatlands in Indonesia by including them in an international cap-and-trade system. In the future, one could include peatland restoration in a greenhouse-gas trading system by giving countries such as Indonesia emission targets that explicitly cover those emissions (which is not the case now). However, the difficulty of setting an appropriate target for such variable emissions should not be underestimated.

If there is a risk that Indonesia will again experience peatland-related emissions similar to those released in the 1997 El Niño event (13–40% of global emissions), it is hard to see how the country could accept a target that would make it accountable for those emissions; reasonably enough, Indonesians consider that they have little control over them.

On the other hand, offering a generous emission target, which would cushion such emission events, runs the risk of inflating the international carbon market with ‘tropical hot air’ if it turns out that the peatlands do not burn. This would have detrimental consequences for emission-reduction efforts in other parts of the world.

Clearly, these problems affect not only Indonesia but also other countries with

large potential changes in their biospheric stock of carbon. Brazil, just like Indonesia, gets a large share of its emissions from deforestation. Annual fluctuations in deforestation-related emissions are as large as the entire Kyoto target for the European Union (see U. M. Persson and C. Azar, *Brazil beyond Kyoto*, Swedish Environmental Protection Agency, Stockholm, 2004).

In addition to problems in setting appropriate targets, there are huge uncertainties in emission estimates and difficulties in separating human-induced emissions from natural emissions.

One possible way forward is to develop a separate protocol for these highly variable emissions, based on specific policies and measures that encourage better land-use patterns and protection of sensitive ecosystems. This would avoid the uncertainties of target-setting altogether.

Alternatively, if all carbon emissions were included under the same cap, the target for such variable emissions could be made non-binding, so that carbon credits could only be claimed if emissions fell below a certain level. This would make it politically feasible to negotiate a reasonably tough target, because countries would not be held accountable for emissions exceeding the target. An incentive for peatland restoration and reduced deforestation would be created, while the risk of creating ‘tropical hot air’ would be diminished.

Finally, although there are difficulties and problems associated with almost all proposals to deal with these problems through international climate negotiations,

this should not prevent us from taking action to deal with it on the ground today.

Christian Azar, U. Martin Persson

*Department of Physical Resource Theory,
Chalmers University of Technology,
S-412 96 Gothenburg, Sweden*

Mouse geneticists need European strategy too

Sir—Your News story “Geneticists prepare for deluge of mutant mice” (*Nature* 432, 541; 2004) promises mutants “covering every single gene”. But the section headed “It's a knockout” notes that many genes will be missed if a simple knockout strategy is exclusively used, owing to embryonic lethal knockouts and compensatory mechanisms.

Clearly, other strategies are needed if the promise to genome researchers is to be kept. The conditional knockout method proposed by a European consortium is intended for use in addition to the simple knockout strategy — not instead of it. Although the conditional knockout method may be more expensive and time-consuming than the simple knockout method, it offers a far more comprehensive picture of gene function, and allows the selection of somatic mutants, which are more closely related to human disease conditions.

It is only through the combination of methods that the News story's promise can become a reality.

Wolfgang Wurst

*Institute of Developmental Genetics,
GSF Research Centre, Ingolstaeder Landstraße 1,
85764 Munich/Neuherberg, Germany*

Contemplating the abyss

The role of environmental degradation in the collapse of human societies.

Collapse: How Societies Choose to Fail or Succeed

by Jared Diamond

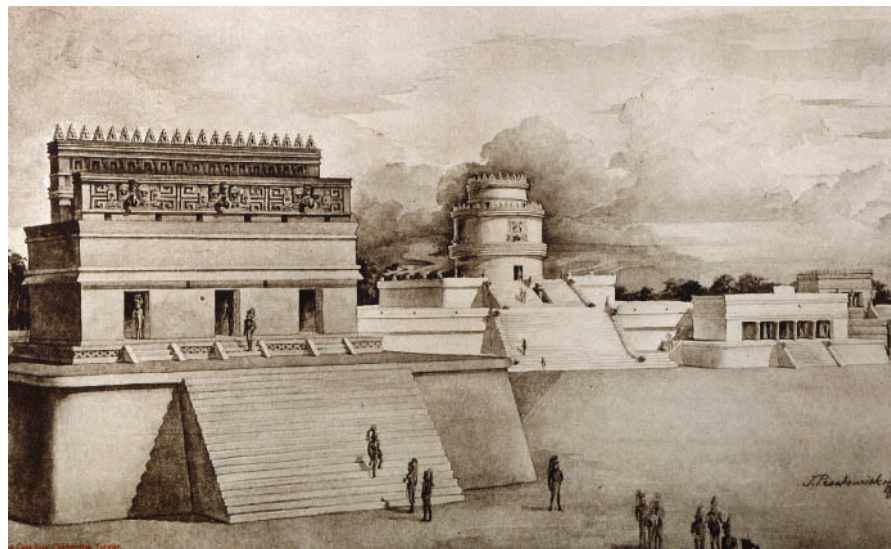
Viking/Allen Lane: 2005. 592 pp. \$29.95/£20

William Rees

Jared Diamond is a necessary antidote to Bjørn Lomborg (of *The Skeptical Environmentalist* fame) and other latter-day acolytes of Julian Simon who preach that environmental problems are largely bogus and that the human future is secure. Consider the facts: industrial humans are the most voracious predators in the world's oceans and, simultaneously, the most successful terrestrial carnivore ever to have walked the Earth. As if to underscore the merits of our generalist evolutionary strategy, we are also the dominant herbivore in grasslands and forests all over the planet (especially if we consider the demands of our 'industrial metabolism'). In short, humans are the most ecologically significant macro-consumers in every major ecosystem type on Earth (with the notable exception of deep marine vent communities, which we have only just begun to explore), and we are literally consuming ecosystems from within. Meanwhile, Earth scientists say that human activity is the most powerful geological force altering the face of the planet, and the erosive pace is accelerating.

Nothing bogus here — this is an ecological reality. Human behaviour towards the ecosphere has become dysfunctional and now arguably threatens our own long-term security. The real problem is that the modern world remains in the sway of a dangerously illusory cultural myth. Like Lomborg, most governments and international agencies seem to believe that the human enterprise is somehow 'decoupling' from the environment, and so is poised for unlimited expansion. Jared Diamond's new book, *Collapse*, confronts this contradiction head-on. It is essential reading for anyone who is unafraid to be disillusioned if it means they can walk into the future with their eyes open.

As suggested by its title, this book is about societal collapses — past, present and future — and the factors that cause human societies to fail. But it is also a history of success, of societies that were able to confront their problems and thrive, sometimes for millennia. Diamond reasons that, for all the trappings of modernity, the human past presages the human future, and thus provides "a rich database from which we can learn". His primary mission is to determine the ecological, political and cultural conditions that lead



Before the fall: Chichén Itzá was a thriving city until the collapse of the Mayan civilization.

to collapse and to contrast these with the conditions that favour success.

Diamond defines collapse as "a drastic decrease in human population size and/or political/economic/social complexity, over a considerable area for an extended time". He founds his analysis on systematic consideration of five sets of causal mechanisms. Any of the first four sets — damage that people inflict on their ecosystems, climate change, the actions of hostile neighbours, and loss of contact with trading partners (friendly neighbours) — may or may not be relevant to any particular case. The fifth set, however — how a society responds to the other classes of problems as they arise — is always a determinant of that society's future.

Collapse is based on a series of detailed case studies. Diamond begins with an affectionate portrait of modern-day Montana, revealing many of the socio-political and environmental uncertainties that cloud the state's future. The main purpose here is actually to establish common themes for subsequent chapters on societies that have long since completed the cycle to collapse: cultures on Easter Island, Pitcairn Island and Henderson Island in the South Pacific; the native American culture of the Anasazi; the Maya; and the Norse Greenland culture. These tragic failures are followed by several uplifting cases of societal success, including Tikopia in the South Pacific, the New Guinea highlands, and Japan during the Tokugawa era.

Diamond then provides a fuller exploration of the many rich parallels between these historic cases and select modern societies. The latter include the contemporary malthusian disasters of Rwanda and Haiti;

the success (by developing-world standards) of the Dominican Republic; an emerging developing-world giant, China, which is the scariest case because of the staggering scale of its problems and potential global impacts; and Australia, a developed-world society reeling from ecological degradation but beginning to respond creatively. (Tellingly, however, Diamond's most realistic scenario for Australia sees it falling into decline under the weight of accelerating environmental problems, perhaps just ahead of the rest of the developed world.) Curiously missing from this section is a detailed consideration of the United States, Diamond's own country and the one imposing the greatest ecological load on the planet.

What emerges most clearly from Diamond's analysis is the central role played by environmental decay in undermining human societies. Eight ecological processes familiar to environmentalists today also plagued earlier societies: habitat destruction (such as deforestation and desertification), soil degradation (erosion, water-logging and salination), water supply problems, over-hunting, over-fishing, the impacts of introduced species, population growth pressures, and rising per capita impacts. The relative significance of each of these processes varies greatly from case to case, but all the ancient societies examined put themselves at risk, sometimes fatally so, by inadvertently undermining the very ecosystems that supported them — and modern societies have even more ecological spectres to banish.

Highlighting environmental degradation as a fundamental factor in societal collapse distinguishes Diamond's interpretation from

that of Joseph Tainter in his 1988 book *The Collapse of Complex Societies*, which has long been the best-known book on the subject. Tainter developed a convincing argument that societies actually advance or 'complexify' as they respond creatively to major challenges. He therefore found it difficult to accept that any complex society with pre-developed administrative, organizational and technical coping skills would allow itself to succumb to emergent ecological problems. Instead, he placed the blame for collapse on socio-political instability resulting from diminishing returns to investment in problem solving — that is, on excessive complexity. Diamond concedes that the implosion of a vulnerable society might be triggered by an overstretched economy, dissolute leadership or enemy invasion, say, but argues that the ultimate cause is usually fragility caused by ecological degradation.

In the book's final section, Diamond focuses on practical lessons. Why do societies sometimes make such disastrous decisions? What can we 'moderns' usefully learn from the responses of ancient societies to environmental crises? What is the appropriate role of the private sector, transnational corporations in particular? Which of today's environmental trends are the most threatening and how do they differ from those that sank previous societies? Anticipating resistance to his findings from perennial optimists, Diamond includes well reasoned ripostes to a dozen common 'one-liner' objections to the seriousness of environmental problems and to the relevance of previous collapses to techno-industrial society.

In the end, Diamond's painstaking toil in the deep mines of history rewards him with sufficient nuggets of hope that he emerges "cautiously optimistic" about the human prospect. Modern society's ecological and geopolitical problems may be daunting but, in theory, they can be solved if we take the right decisions to reduce our ecological footprints. And let's not forget that we are uniquely positioned to learn from the collapse of previous societies.

Regrettably, theory and example do not always translate into practice. The most important lesson to be drawn from *Collapse* is that resilient societies are nimble ones, capable of long-term planning and of abandoning deeply entrenched but ultimately destructive core values and beliefs. This, in turn, requires a well informed public, inspired leadership and the political will to take decisions that go against the established order of things. In this light, the astute observer of contemporary geopolitics and ecological decline might be excused a descent into quiet despair. ■

William Rees is professor of ecological planning in the School of Community and Regional Planning, University of British Columbia, 6333 Memorial Road, Vancouver V6T 1Z2, Canada.

A natural pioneer

John James Audubon: The Making of an American

by Richard Rhodes

Knopf, 2004. 528 pp. \$30

John Fitzpatrick

Some of the best-known icons of America's pioneer age endure through bloated mythologies of heroism generated by salesmanship and an American culture perpetually in demand of larger-than-life heroes. Not so John James Audubon, the most famous naturalist-artist in history. His monumental double-elfant folio production *Birds of America* holds a deserved place among the greatest artistic achievements of any era, and his writings provide our richest chronicle of North American natural history before the industrial revolution.

The mythology surrounding Audubon's life and character bestows on him a mixed review: mysterious birth and early childhood, commercial failure and debtors' prison, and long absences from his family. Audubon the legend has been painted as a bumbling businessman, a self-promoting dandy and a gallivanting nature-boy who abandoned wife and children for long periods to live in the woods and sketch birds. Audubon's serendipitous good fortune in marrying a patient and hard-working wife, Lucy Bakewell, and later becoming a partner with virtuoso engravers the Robert Havells, senior and junior, carry as much weight in fable as do his keen observational skills, passionate creativity, breakthrough artistic genius and scientific contributions.

In this masterful biography, the historian and novelist Richard Rhodes systematically

debunks these persistent canards. This is a lively, interpretative tour through the personality, travels, business ventures, accomplishments and foibles of an extraordinary man. Beginning with Audubon's illegitimate birth in Saint-Domingue (now Haiti) in 1785, Rhodes weaves together each turn of Audubon's difficult but colourful 61-year life with insightful commentary and vignettes that illustrate what life was like in the forested frontier west of the Appalachians. He underscores the turbulent social and economic forces at work in both Europe and America during the early nineteenth century.

Supported by generous quotes from his copious journals and correspondence, the Audubon we see here was a multi-talented, complex, hard-working and genuinely heroic family man. Handsome and lively — "his eyes alone commanded attention" — he had unrivalled physical dexterity and energy, loved to sing and dance, played the violin and flute, and excelled at fencing, marksmanship, horsemanship and — of course — art. Everyone he met remembered him vividly, and Rhodes' intimate portrait leaves little doubt that the reader would love to have known this man personally.

Rhodes casts Audubon's most famous business failures as earnest and promising ventures that collapsed at precisely the same moments, and for the same reasons, as virtually every other similar business in the country. After a shipping partnership in New Orleans was destroyed by embargoes caused by the war of 1812, Audubon successfully built up a new one along the rapidly developing Ohio River. Then, during America's first major economic depression (the panic of 1819), nearly every mercantile business in the country collapsed. Returning home after failing to collect a debt in New Orleans,



An eye for detail: this image of long-billed curlew is typical of Audubon's bird paintings.

Audubon “took an upriver Mississippi steamboat to the mouth of the Ohio and walked the last 130 miles home”.

The heart of the story begins here, as the 32-year-old Audubon plunges from prosperous Ohio gentleman to a penniless, depressed woodsman with a growing family to feed. With understated epiphany — “nothing was left to me but my humble talents” — he soon excelled at drawing portraits of wealthy citizens, while keeping alive his obsessive passion for birds. Long before imagining how he would use them, Audubon had been creating a portfolio of masterpieces. His life-sized paintings of birds were not the stiff scientific illustrations then in fashion. A keen observer, Audubon committed himself early to capturing (often by exaggeration) each bird’s species-specific personality. In 1810, Alexander Wilson, the ‘father of American ornithology’ but a depressed loner, encountered Audubon in Louisville. Jealous of the young artist’s superior talent, Wilson refused Audubon’s offer to collaborate, and died just three years later. But Wilson’s big idea — to travel about America painting and writing about its birds — left a lasting impact. As he began to paint for a living, Audubon realized his calling and single-mindedly pursued Wilson’s idea. By 1926 he was sailing for England to find an engraver and begin publication, and he would soon leave Wilson in the dust.

Rhodes traces the artist’s meteoric rise as England embraced both the art and the man. Audubon’s revolutionary paintings portrayed highly animated birds in exacting detail, reflecting frontier America in vivid, even bloody, colour. His detailed knowledge about the lives of mostly unfamiliar birds impressed England’s stuffy scientific circles. Audubon’s dogged pursuit of a one-man business demanded a long and punishing schedule. Rhodes gives haunting, nuanced colour to the picture of Audubon in England, struggling to gain credibility and subscribers. Steadily achieving fame, the artist is wracked with depression, self-doubt, changes of plans and sadness. We see him as a passionate, profoundly tender man who deeply misses his wife. Anyone who thinks they know the travails of a relationship at a distance should read what these two lovers endured, at a time when their frequent letters to each other either disappeared or took six months to be delivered.

I found it only mildly disappointing that Audubon’s scientific relationships and contributions are treated more lightly than his personal and business affairs. We are given glimpses, for example, of his election as a fellow of the Royal Society of London (only the second American, after Benjamin Franklin), his brief association with William Swainson, his long friendship with Charles Bonaparte, his enmity with George Ord and the Philadelphia establishment, and his close partnership with naturalist John Bachman

Science in culture

What’s in the flask?

The origin of the archetypal image of the chemist.

Philip Ball

What are these scientists all looking at? The archetypal image of the chemist, ubiquitous in stock photographic images today, and even in clip-art databases, depicts a lab-coated figure gazing at a flask of liquid held aloft. The inclusion of the picture in the bottom right will be understood by British readers, who may recognize the features of a woman who went on to become the country’s prime minister.

But this is not what real chemists spend their time doing. So where does the pose come from? As Joachim Schummer and Tami Specter pointed out at a recent conference in Paris on the public image of chemistry, the answer lies in the image in the top left. This appeared in a book dating from 1283, the Latin translation of Avicenna’s *Canon of Medicine*, and shows not a chemist but a doctor. The flask contains not a solution synthesized by alchemy but a sample of a patient’s urine — diagnoses were typically made by uroscopy, the practice of inspecting the urine for colour, clarity and other qualities.

When Paracelsus introduced chemistry into medicine (so-called iatrochemistry) in the early seventeenth century, this image of the gazed-at flask transferred itself from medicine to ‘chymistry’, and subsequently became so much a part of the subject’s visual language that it is alive and well today.

Philip Ball is a consultant editor for Nature.

► www.hyle.org/service/chmc2004/



The gazed-at flask (clockwise from left): Avicenna’s doctor; Gerrit Dou’s *The Urine Doctor*; a modern chemist; Margaret Thatcher; the ClipArt view.



of Charleston. But we barely meet William MacGillivray of Edinburgh, with whom Audubon wrote the five-volume *Ornithological Biography*, his most important and lasting scientific achievement.

The book contains numerous errors of nomenclature, and would have benefited from proofreading by an ornithologist. Most disappointing of all to me are the illustrations, which are mostly small black-and-white pictures, many untitled. The printing quality of the 16 colour plates is abysmal — Audubon would never have approved them for public release.

Rhodes has significantly clarified both the factual record and the human understanding

of this truly legendary man whose name has become synonymous with birds. Of French descent and English fame, Audubon became a consummate American who realized that his work would become immortal as his beloved frontier began disappearing. Indeed, with each passing year, Audubon’s legacy continues to multiply in value, and this book will add immeasurably to the world’s deep appreciation for his passion.

John Fitzpatrick is at the Cornell Laboratory of Ornithology, Ithaca, New York 14850, USA.

More on Audubon

Under a Wild Sky by William Souder
North Point Press: 2004. \$25.

Bridging the gap

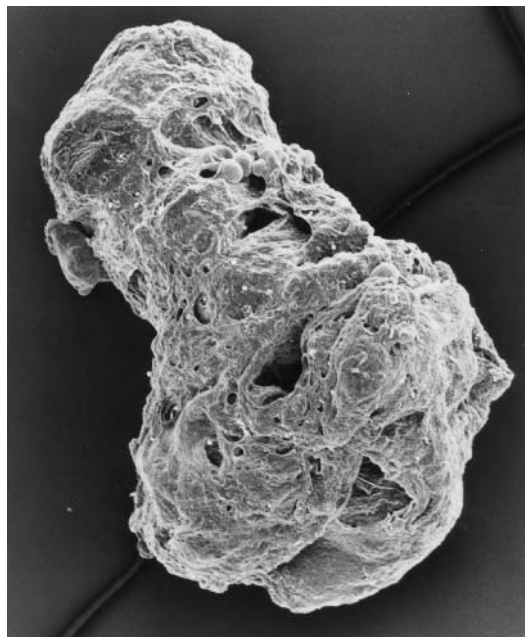
Tissue engineering: mathematical models are helping to take tissue engineering from concept to reality.

Ben D. MacArthur and
Richard O. C. Oreffo

The emerging discipline of tissue engineering has the grand aim of understanding the principles of tissue growth, and applying this to produce functional replacement tissue for clinical use. There have been several remarkable successes. A recent example is the work of Patrick Warnke, Hendrick Terheyden and co-workers, in which a section of replacement jaw was generated inside a sculpted titanium mesh cage by coaxing appropriate stem cells (those elusive precursor cells that give rise to specialized body tissues) to form bone. But although successes such as this show that, in concept, tissue engineering is possible, routine implementation of such strategies remains some time off.

In fact, this is no big surprise, as such a broad implementation requires a much better understanding of the principles of tissue formation than we currently possess, from the fundamentals of stem-cell biology to the physics and biomechanics of pattern formation. To complicate (or perhaps, enrich) matters further, this mandate also calls on the specialist expertise of scientists from a wide variety of disciplines — such as cell and molecular biologists, clinicians and materials scientists — each of whom sees the various problems involved from the perspective of their own discipline. Practical integration of these seemingly disparate strands of understanding is proving to be a rich source of scientific challenge and opportunity. In particular, collaborations between biologists and mathematicians are now providing alternative and often innovative ways of thinking about tissue regeneration.

Reproducing functional tissue *ex vivo* requires an understanding not only of the behaviour of individual cells, but also of how global form and function arise from local cellular interactions. By looking at evolving tissue as a complex biological system, mathematical models can provide just such a holistic understanding. The use of agent-based models to interpret stem-cell systems is beginning to show promise in offering new ways of thinking about tissue evolution. In these models, cells are considered as distinct entities (or agents) positioned on an appropriate lattice, and simple cellular behaviours are prescribed, which, on their own or on the local scale, are insufficient to produce pattern. But on the global scale, structure is seen to emerge from long-range summation of these



Growing bone (above) on a scaffold is still far from routine.

low-level behaviours. Such models are now being incorporated into practical work programmes to explore the behaviour of stem-cell systems and mechanisms of tissue regulation.

As a related example, our current work focuses on the behaviour of selected stem-cell populations *in situ*, as they progress through the osteogenic route to form bone. Behaviour here includes both the differentiation potential and the spatio-temporal patterns of adhesion, migration and proliferation of the cells. In particular, we are using mathematical models of cell-population behaviour in conjunction with experimentation to explore regulation of the osteoblast and bone-tissue phenotypes on various three-dimensional porous scaffolds. By combining expertise in biomimetic materials science and stem-cell biology with mathematical models, our aim is to select the tissue-engineering strategies that are most likely to be successful and offer creative ways of investigating tissue formation. This work is directing new experimental research that is helping to elucidate relationships between stem-cell activity, differentiation, nutrient delivery and evolving macroscopic tissue architecture.

A final and appealing new direction is the use of complex network theory to analyse the 'shape' of phenotypic regulatory mechanisms. In these models, topologically complex networks — consisting of all potential regulators of a cellular phenotype and their interactions — may be generated from the

wealth of stoichiometric data that are becoming increasingly available. As there may be hundreds of factors implicated in a given phenotype, such networks may be bewilderingly entangled and appear intractable at first sight. But statistical comparison of such 'real-life' networks with properties of similar-sized random networks (those in which network topology is generated, in some prescribed sense, randomly) can begin to relate the shape of the real-life network to its function. For example, the features of the real-life network that occur significantly more often than in the random networks may be identified for further investigation, as may those vertices, edges or subnetworks whose presence or absence profoundly affect the global properties of the network. Such critical subnetworks may represent functionally significant control motifs, whereas

the elements whose removal has little effect on global network properties may be considered to be more functionally peripheral. Thus, representing phenotypic regulatory mechanisms as complex networks may allow the fundamental functional units of morphogenesis to be redefined in terms of, for example, small networks of genes, transcription factors and proteins, rather than in terms of these elements in isolation.

These are bright times for tissue engineering. The integration of mathematical modelling with experimentation in an iterative framework — each informing and directing the other — is offering exciting challenges, as well as substantial scope to further our understanding of tissue regeneration. In the end, this may prove crucial in taking tissue engineering from concept to reality. ■

Ben D. MacArthur and Richard O. C. Oreffo are in the Bone and Joint Research Group, Division of Developmental Origins of Health and Disease, University of Southampton, Southampton General Hospital, Southampton SO16 6YD, UK.

FURTHER READING

Lanza, R. P., Langer, R. & Vacanti, J. P. (eds) *Principles of Tissue Engineering* (Academic, New York, 2000).
Lauffenburger, D. A. & Griffith, L. G. *Proc. Natl Acad. Sci. USA* **98**, 4282–4284 (2001).
Murray, J. D. *Mathematical Biology II: Spatial Models and Biomedical Applications* (Springer, Berlin, 2003).
Newman, M. E. J. *SIAM Rev.* **45**, 167–256 (2003).
Warnke, P. H. *et al. Lancet* **364**, 766–770 (2004).

Silver nanoswitch

Jan van Ruitenbeek

Ionic conductors have many applications — in sensors, fuel cells and batteries. Are nanoelectronic devices based on ionic conductors now about to replace silicon?

Most electronic appliances are based on digital electronics, which in essence just require a lot of switches working together in an organized fashion. Much research has been aimed at finding a reliable switching mechanism that can beat conventional silicon technology to permit ever smaller and more powerful electronics. The ideal switch should be scalable down to atomic size; it should have low power consumption, and require just two leads for both read and write memory operations. On page 47 of this issue¹, Terabe and co-workers describe an invention that comes close to this ideal. They exploit the fascinating properties of silver sulphide, a material in which electrical conductivity is carried by both electrons and silver ions. The resulting devices can be used for logic as well as for fast memory operations, and they function at room temperature.

In most solids, atoms sit at fixed positions in a regular crystal lattice. In the solid ionic conductors used by Terabe *et al.*, however, some ions have many possible equivalent positions in the lattice and can wander through the material. Figure 1 illustrates this for the conductor of interest here, Ag_2S . When the material is connected by two silver leads to a battery, Ag^+ ions are formed at the interface between silver sulphide and the positive silver electrode, while Ag^+ is reduced at the other electrode. This process

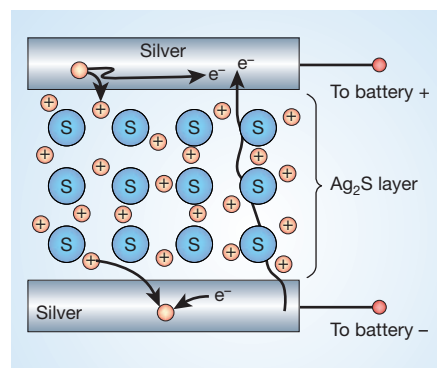


Figure 1 Silver sulphide — a mixed electronic and ionic conductor. Two silver contacts, at the top and bottom, are applied to Ag_2S and connected to a battery. The current is partly carried by electrons, partly by positive silver ions (circled 'plus' signs) diffusing through the sulphide in the opposite direction. The ions are replenished at the positive electrode by oxidation of the electrode material, while silver is reduced and deposited at the other end.

leads to the transport of silver, removing it from the positive lead and depositing the same quantity at the negative lead. Ag_2S is one of a rare kind of solid ionic conductors that have two unusual features: it operates at room temperature and it conducts electrons as well as ions. Both features are of central importance to the device created by Terabe and co-workers.

A few years ago, the authors reported that nanoscale silver mounds formed on top of a Ag_2S crystal when a scanning tunneling microscope (STM) was used^{2–4}. In that experiment, a silver bottom electrode is held at a positive electrical potential with respect to the platinum STM tip. Electrons tunnelling from the tip to the surface of Ag_2S are partly used up in reducing Ag^+ ions to metallic silver. Keeping the tip at a fixed height above the surface results in the formation of a silver metallic bridge between tip and sample. The process can be reversed by reversing the electrical potential, which dissolves the silver bridge back into the sulphide. This is the principle of the switch: contact can be made or broken by applying a voltage of the appropriate sign. The reason this work went largely unnoticed is that many switching mechanisms between STM tips and substrates have been discovered in recent years, but they are of little practical value because each device requires its own STM. For practical applications, such tunnel junctions between two electrodes need to be controlled in a simpler and more reproducible way.

Terabe *et al.*¹ have come up with a clever solution: they exploited the properties of the ionic conductors themselves to create and control the required tunnel gap. A layer of Ag_2S on top of a silver wire is in contact with a thick platinum wire through a silver layer one nanometre thick (Fig. 2a). The platinum and silver leads are then connected to a voltage source to run an electronic current from top to bottom. This current is accompanied by the transport of silver downwards through the silver sulphide, and after a few seconds the top silver layer vanishes, resulting in a loss of contact with the platinum lead. The device is now in the 'off' state and ready for operation. When the polarity of the applied voltage is reversed (Fig. 2b), a local silver bridge is promptly formed which again closes the gap between the platinum and Ag_2S , turning the switch 'on'. The process can be reversed and repeated rapidly because only a few atoms are involved.

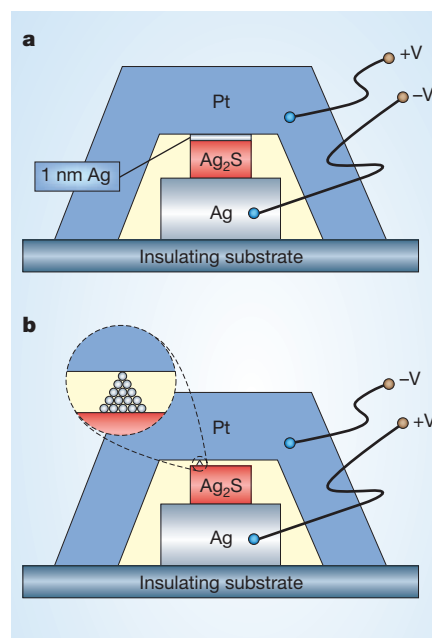


Figure 2 A rewritable memory bit based on the properties of the silver sulphide (Ag_2S) mixed ionic conductor, as described by Terabe and colleagues¹. a, A one-nanometre-thick silver layer deposited on top of the Ag_2S layer disappears into the sulphide layer when a current flows from the platinum (Pt) lead to the silver lead. This results in loss of contact between the two electrodes and initializes the device. b, A bridge of silver atoms is locally formed by applying a voltage of opposite sign, re-establishing contact between the silver sulphide and platinum. The conductance through the device can be as small as one quantum unit of conductance, suggesting that the silver bridge can touch the platinum lead with just one atom.

Terabe *et al.* observe, moreover, that the conductance of the device can be as small as one quantum unit if a short voltage pulse of the correct amplitude and duration is applied. In this case, it seems that the silver bridge has grown upwards until just one atom touches the platinum lead (for a review, see ref. 5). To switch between on and off states requires voltages higher than 100 mV. The state of the memory bit — that is, whether on or off — can be read non-destructively at voltages lower than that, taking advantage of the electron-conducting property of the silver sulphide.

By combining two silver sulphide switches with resistors and capacitors, Terabe *et al.*

carry out the basic logic operations AND, OR and NOT. In principle, this is all that is needed to perform more complex logic operations. However, the efficiency of these logic gates will be significantly reduced when their inputs and outputs are connected to other logic gates in a large digital circuit. In order to avoid this problem, there should be some way of amplifying the gate signals; otherwise, their logic applications will unfortunately be limited. The multiple steps in quantum conductance that the authors observe with increasing voltage are also remarkable, but probably not of sufficiently practical use in view of their limited reproducibility.

Yet the main result is of great beauty and

simplicity, and is scalable to nanometre-sized addressable bits. The authors have done well to protect their work with several patent applications. ■

Jan van Ruitenbeek is in the Kamerlingh Onnes Laboratory, Leiden University, Niels Bohrweg 2, Leiden 2333 CA, The Netherlands.
e-mail: ruitenbeek@physics.leidenuniv.nl

1. Terabe, K., Hasegawa, T., Nakayama, T. & Aono, M. *Nature* **433**, 47–50 (2005).
2. Terabe, K., Hasegawa, T., Nakayama, T. & Aono, M. *RIKEN Rev.* **37**, 7–8 (2000).
3. Terabe, K., Hasegawa, T., Nakayama, T. & Aono, M. *Appl. Phys. Lett.* **80**, 4009–4011 (2002).
4. Terabe, K., Hasegawa, T., Nakayama, T. & Aono, M. *J. Appl. Phys.* **91**, 10110–10114 (2002).
5. Agrait, N., Levy Yeyati, A. & van Ruitenbeek, J. M. *Phys. Rep.* **377**, 81–279 (2003).

Cognitive science

Staring fear in the face

Patrik Vuilleumier

The unusual case of SM, a person who has a very specific deficit in recognizing fearful expressions on people's faces, is providing intriguing insights into how we perceive emotion.

Charles Darwin thought that the ability of humans to display and perceive emotional states on a face evolved to convey non-verbal signals rapidly¹. If an individual's expression could communicate a potential threat, for example, his neighbours would be able to respond quickly and direct their attention to the source of the danger. Thus, a common view is that the perception of fear might guide appropriate visuomotor behaviour². In a striking reversal of this perspective, work by Adolphs *et al.* on page 68 of this issue³ suggests that discerning fear in faces may depend on how one scrutinizes them in the first place.

The authors describe a patient (SM) who has bilateral brain lesions in the amygdala, a region of the medial temporal lobe known to be critical for the perception of fear⁴. SM cannot recognize fear from facial expressions⁵, and Adolphs *et al.* show that this is because she fails to look spontaneously towards the eyes on a face. When shown a face displaying an unmistakable expression of terror, she tends to fixate unworriedly on the nose and mouth regions, neglecting to notice the wide, scared eyes. Thus, she erroneously judges that the face has a neutral expression. By contrast, normal people always look immediately at the eye region of a face, and all the more so when the face is fearful⁶.

SM avoids the eyes of all faces, no matter what their expression. But, remarkably, only her perception of fear is impaired — she can recognize other emotions. This suggests that visual cues provided by the eyes are particularly critical for the recognition of fear; other facial emotions can presumably be recognized without looking at the eyes (happiness

can be inferred from a smile, for example). SM was also tested on a 'bubble' visual task⁷, in which she had to discriminate between fearful and happy faces seen through apertures that revealed only small parts of the image. This allowed the investigators to determine which region of the face she used to distinguish the expressions. Again unlike normal individuals, SM failed to use information from the eye area, but she could still take cues from around the mouth.

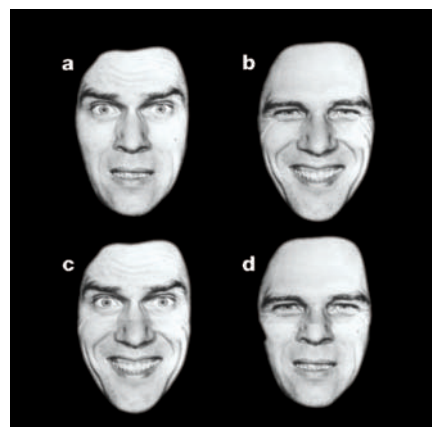


Figure 1 How the eyes contribute to facial expressions of fear. a, b, Examples of fearful (a) and happy (b) faces from a standard image set¹⁹. c, d, Composites using the top and bottom halves from the same faces with fearful eyes but happy mouth (c) or happy eyes but fearful mouth (d). As can be seen, the global configuration in composites is affecting the perceived emotion expressed by the face. A study by Adolphs *et al.*³ sheds new light on how the amygdala in the brain is involved in processing the eyes in such expressions.

Most surprisingly, simply instructing SM to 'look at the eyes' could restore normal recognition of fearful expressions, indicating that she still knows what fear 'looks like' but seems unable to notice scared eyes when she is not prompted to look at them. This 'rescue' was short-lived, however, and SM needed to be reminded continually to look at the eyes. These new results unexpectedly reveal that the damage to the amygdala might impair attention and exploration strategies, rather than causing a perceptual deficit affecting the visual analysis or categorization of specific facial traits.

Much recent research has focused on the role of the human amygdala in fear recognition. Numerous brain-imaging studies confirm that the human amygdala responds more to fearful faces than to faces expressing other emotions, but the exact function of the amygdala during recognition of facial expressions remains a mystery. Initially, the observation that SM's perception of fear is impaired while her recognition of other emotions remains intact⁵ was thought to support the idea that different categories of emotion involve distinct neural circuits in the brain⁸. The findings of Adolphs *et al.*³ now suggest a very different mechanism, perhaps involving a more general role for the amygdala in modulating visual and attentional processing⁹. The amygdala is known to be sensitive to perceived gaze direction, responding most when the eyes in a facial image seem to be looking at the observer¹⁰. In agreement with Darwin's theory, it makes sense if fear perception is intimately connected with locating the threat that fearful eyes are seeing¹¹. The simplicity of such a mechanism might allow for swift responses to danger, even with poor or crude inputs, or during inattention. Indeed, it was recently found that when a normal subject is shown shapes that look like the whites of a pair of eyes, his amygdala responds more to larger shapes (corresponding to wide, fearful eyes) than to small (happy) shapes¹².

However, the amygdala is probably not just an 'eye detector', and perception of fearful expressions is unlikely to rely solely on wide eyes. Previous research^{13–15} suggests that processing single 'diagnostic' features in faces is not sufficient to appraise their expression fully, but that more global configural information is important (for example, see the composite faces in Fig. 1). Moreover, the bubble task might induce a bias to use the local details visible through the bubble apertures rather than configural information, which would be more natural¹⁶, particularly in a dichotomous fearful–happy classification task (for instance, SM might simply check for the presence of a smile, and therefore never need to look at the eyes to perform this particular task). The demands of particular tasks also influence whether the subject uses local or global visual features

during face processing¹⁷. Furthermore, brain-imaging data indicate that even though the amygdala might respond to fearful eyes when they are presented alone, it is activated most in response to whole faces¹⁸.

Finally, it remains to be determined whether SM's attention to other facial features is normal (only her response to the eye region was recorded), and to explain why she can still recognize expressions of sadness or anger in which eye information is important (normal subjects find it more difficult to recognize these emotions when the eyes are erased)³.

The intriguing implications of these new findings need to be explored. What are the neural circuits by which the amygdala might guide eye scan-paths? How does SM judge expressions in composite faces such as those in Figure 1? How does she perform on more implicit tests of fear recognition, or using graded rather than dichotomous measures? Does she orient her eyes normally to emotional visual stimuli other than faces, and to emotional voices? What is the amygdala's normal role in exploring social situations and looking at other people, and are these mechanisms altered in diseases such as phobias or autism that are thought to involve the amygdala? We are just beginning to realize how the brain processes emotionally

relevant cues in the environment, and the unusual features of SM will provide much food for future thought.

Patrik Vuilleumier is in the Laboratory for Behavioural Neurology and Imaging of Cognition, Clinic of Neurology and Department of Neurosciences, University Medical Center, 1 rue Michel-Servet, 1211 Geneva 4, Switzerland. e-mail: patrik.vuilleumier@medecine.unige.ch

1. Darwin, C. *The Expression of Emotion in Man and Animals* (Oxford Univ. Press, 1872).
2. Ohman, A. *Psychophysiology* **23**, 123–145 (1986).
3. Adolphs, R. et al. *Nature* **433**, 68–72 (2005).
4. LeDoux, J. E. *Annu. Rev. Neurosci.* **23**, 155–184 (2000).
5. Adolphs, R., Tranel, D., Damasio, H. & Damasio, A. *Nature* **372**, 669–672 (1994).
6. Yarbus, A. L. *Eye Movement and Vision* (Plenum, New York, 1967).
7. Gosselin, F. & Schyns, P. G. *Vision Res.* **41**, 2261–2271 (2001).
8. Calder, A., Lawrence, A. & Young, A. *Nature Rev. Neurosci.* **2**, 352–363 (2001).
9. Vuilleumier, P., Richardson, M., Armony, J., Driver, J. & Dolan, R. J. *Nature Neurosci.* **7**, 1271–1278 (2004).
10. Kawashima, R. et al. *Brain* **122**, 779–783 (1999).
11. Sander, D., Grafman, J. & Zalla, T. *Rev. Neurosci.* **14**, 303–316 (2003).
12. Whalen, P. J. et al. *Science* **306**, 2061 (2004).
13. McKelvie, S. J. *Br. J. Social Psychol.* **34**, 325–334 (1995).
14. Calder, A. J., Young, A. W., Keane, J. & Dean, M. J. *Exp. Psychol. Hum. Percept. Perform.* **26**, 527–551 (2000).
15. Prkachin, G. C. *Br. J. Psychol.* **94**, 45–62 (2003).
16. Murray, R. F. & Gold, J. M. *Vision Res.* **44**, 461–470 (2004).
17. Schyns, P. G. & Oliva, A. *Cognition* **69**, 243–265 (1999).
18. Morris, J. S., deBonis, M., Dolan, R. J. *Neuroimage* **17**, 214–222 (2002).
19. Ekman, P. & Friesen, W. *Pictures of Facial Affect* (Consulting Psychologists Press, Palo Alto, 1976).

Evolutionary genetics

Differentiation by dispersal

David W. Coltman

Gene flow between populations — caused by migration, for instance — is most often viewed as a homogenizing force in evolution. But two studies of wild birds and non-random dispersal find otherwise.

Whether or not two separate populations of a species become genetically different is thought to depend largely on gene flow. Classical population-genetics theory predicts that populations that frequently exchange individuals through dispersal will remain genetically similar¹. Disconnected populations, by contrast, have a greater capacity to become distinct through forces such as genetic drift and adaptation to local conditions. In population genetics, dispersal is often viewed as a diffusion-like, random process, and selection and genetic variation are assumed to be locally homogeneous. Populations of organisms with high rates of dispersal — such as songbirds — are therefore expected to be fairly genetically alike at small spatial scales. But two new independent studies of wild great tits, *Parus major*, challenge this assumption: they show that when dispersal is non-random, genetic differentiation can be produced at surprisingly fine spatial scales (see pages 60 and 65 of this issue^{2,3}).

Postma and van Noordwijk³ studied clutch size in great tits (Fig. 1) on the tiny — 4,022-hectare — island of Vlieland in the Netherlands from 1975 to 1995. They first found that birds that bred in the western part of the island laid, on average, 1.15 more eggs than birds from the eastern part. How much of this difference is determined by the environment, and how much is genetically controlled? Fortunately, 10% of the females born on one side of Vlieland disperse to breed on the other, and this allowed genetic and environmental effects to be teased apart. The authors' analysis showed that birds of eastern ancestry produced consistently smaller clutches in either environment — so there is clearly a large genetic component to the difference in clutch size between the regions. In fact, genetic effects accounted for about 40% of this difference. But, given that the western and eastern regions are separated by only a few kilometres, and they exchange migrants and receive immigrants from outside Vlieland, why does this genetic difference persist?



100 YEARS AGO

Writing on the subject of "Greek at Oxford," a correspondent of the *Times* again expressed the common belief that "Darwin regretted not having learnt Greek." A letter from Mr. Francis Darwin in the *Times* of December 29, 1904, shows that the statement is altogether opposed to Darwin's views. Darwin says of his education at Shrewsbury School:—"Nothing could have been worse for the development of my mind than Dr. Butler's school, as it was strictly classical, nothing else being taught, except a little ancient geography and history" ("Life and Letters," i., 31). He was, in fact, a victim of that "premature specialisation" which is generally referred to in a somewhat one-sided spirit, and from which the public schoolboy is not yet freed. Mr. Darwin adds:—"If the name of Charles Darwin is to be brought into this controversy it must not be used for compulsory Greek, but against it. In 1867 he wrote to Farrar, 'I am one of the root and branch men, and would leave classics to be learnt by those alone who have sufficient zeal and the high taste requisite for their appreciation' ('More Letters of Charles Darwin,' ii., 441)."

From *Nature* 5 January 1905.

50 YEARS AGO

The expedition organized jointly by the Zoological Society of London and the British Broadcasting Corporation returned to Britain just before Christmas from ten weeks field-work in Sierra Leone, bringing a large collection of animals and a considerable quantity of cinematograph films and sound recordings... One of the main objects of the expedition was to find the nesting habitat of *Picathartes gymnocephala*, a rare passerine bird the systematic position of which is obscure; this bird has seldom been seen alive by Europeans. The habitat was found in difficult hilly bush country, and in spite of the dense shade cast by the forest successful films were made of the birds on and near the nests, of the eggs and of the parents feeding the young by regurgitation. Sound records were also obtained of the voices of the birds in their natural surroundings, and a living specimen was captured and brought to London. Another species never before exhibited in captivity that was successfully sought and found is the brilliantly iridescent emerald starling *Coccycolius iris*.

From *Nature* 8 January 1955.



Figure 1 The great tit: challenging assumptions about gene flow and genetic differentiation.

To answer this, Postma and van Noordwijk examined the viability and fecundity of birds born in the east or west that breed in the other region, and precisely quantified levels of immigration from outside Vlieland. Immigrants and birds born in the west tended to have larger clutches than birds born in the east, regardless of where they bred. However, female birds born in the east seemed to be better adapted to life on Vlieland, because they were twice as likely to survive as birds born elsewhere (perhaps allowing them to have smaller clutch sizes). So, from the standpoint of clutch size and survival, immigrants seemed most closely related to the birds born in the west. Most interestingly, 43% of first-time breeders in the west were immigrants to Vlieland, compared with only 13% in the east.

There has thus been an influx of genes for relatively large clutches, and the higher rate of immigration to the west has resulted in bigger clutches there. Birds in the east have maintained their locally adapted smaller clutch sizes against an influx of 13% immigration. But the west is swamped by immigrant genotypes. So differing levels of gene flow have maintained large genetic differences at a very fine spatial scale.

Postma and van Noordwijk suggest³ that differentiation at this scale because of processes such as these may not be rare (as refs 4 and 5 also suggest), but may remain undetected in other populations because few researchers have appropriate long-term data. Results from a 36-year study of nestling body mass in the same species living in woodlands at Wytham in Oxfordshire, UK, would seem to support this suggestion.

Garant *et al.*² found that the mean mass of birds in the eastern block of Wytham woods

has decreased since 1965, whereas that of birds in the northern block has remained constant. Quantitative genetic analyses demonstrated a genetic component to these trends, and overall there was 50% more genetic variance in the northern population than in the eastern population. These regions are separated by only about 2 kilometres, and on average are both composed of more than 50% immigrants from other parts of the woods or from outside. How has this differentiation persisted under such high rates of potentially homogenizing gene flow?

Heterogeneous dispersal again seems to be the key. Birds from central regions of Wytham woods and immigrants showed non-random dispersal patterns with respect to their weight. Emigrants to the north tended to be larger over time, whereas emigrants to the east tended to be smaller. Even within families, heavier offspring showed an increased tendency to settle in the north.

So, a markedly non-random settlement pattern drives fine-scale genetic differentiation in Wytham woods. But why is there this non-random settlement? Garant *et al.* suggest that temporal and spatial variations in bird density provide the mechanism. Local density has been, on average, twice as high in the east as in the north, producing greater pressures on habitat in the east. However, density has been increasing in the north, as birds there are more likely to survive and be reproductively successful. Taken together, the results suggest that individuals that are genetically predisposed to be larger and heavier have preferentially settled in the lower-density habitat in the north, and this has driven population differentiation.

These studies^{2,3} document remarkable — and quantitatively similar — levels of genetic

differentiation at a very fine scale relative to the birds' dispersal capability. In both cases, differentiation is maintained primarily by non-random dispersal and settlement, and in the absence of major spatial differences in selection. The findings shed new light on the mechanisms of microevolution, because there is no reason to suspect that these phenomena are unique to great tits^{4,5}. But the ability to detect such phenomena depends on the availability of long-term data from continuously monitored populations of marked individuals, and on the existence of biotic or abiotic gradients that may drive the underlying microevolutionary processes. In both of these studies there was sufficient pedigree information to be able to compare the performance of individuals with common genetic backgrounds but inhabiting different environments.

With the increasing use of quantitative genetics analyses such as these^{2,3} in evolutionary studies of wild animals⁶, we ought to be paying more attention to how quantitative genetic variation is spatially and temporally structured. For example, studies that integrate fine-scaled maps of habitat quality with complex pedigrees can look at how genes are distributed spatially, and how they may interact with the environment.

The great advantage of the quantitative genetic approach is that microevolution can be studied in the wild without knowing exactly which genes are responsible for the variation in the physical traits in question. But this is also a great disadvantage, because it would be ideal to see evidence for spatially associated differences in the genetic loci involved⁷. Indeed, a truly mechanistic understanding of microevolution requires an understanding of genetic architecture (the properties of the individual genes underlying variation)⁸. One way of gaining such an understanding of microevolution in nature will be to apply genomics^{9,10} to ecological and evolutionary studies in non-model species, using comparative approaches^{11–13}. ■

David W. Coltman is in the Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada.
e-mail: david.coltman@ualberta.ca

1. Slatkin, M. *Science* **236**, 787–792 (1987).
2. Garant, D., Kruuk, L. E. B., Wilkin, T. A., McCleery, R. H. & Sheldon, B. C. *Nature* **433**, 60–65 (2005).
3. Postma, E. & van Noordwijk, A. J. *Nature* **433**, 65–68 (2005).
4. Blondel, J., Dias, P. C., Ferret, P., Maistre, M. & Lambrechts, M. M. *Science* **285**, 1399–1402 (1999).
5. Chan, Y. & Arcese, P. J. *Evol. Biol.* **16**, 939–947 (2003).
6. Kruuk, L. E. B. *Phil. Trans. R. Soc. Lond. B* **359**, 873–890 (2004).
7. Purugganan, M. & Gibson, G. *Mol. Ecol.* **12**, 1109–1112 (2003).
8. Mackay, T. F. C. *Annu. Rev. Genet.* **35**, 303–339 (2001).
9. Feder, M. E. & Mitchell-Olds, T. *Nature Rev. Genet.* **4**, 651–657 (2003).
10. Thomas, M. A. & Klaper, R. *Trends Ecol. Evol.* **19**, 439–445 (2004).
11. Primmer, C. R., Borge, T., Lindell, J. & Saetre, G. P. *Mol. Ecol.* **11**, 603–612 (2002).
12. O'Brien, S. J. *et al. Science* **286**, 458–481 (1999).
13. Aitken, N., Smith, S., Schwarz, C. & Morin, P. A. *Mol. Ecol.* **13**, 1423–1431 (2004).

Geochemistry

Neon illuminates the mantle

David W. Graham

The outer Earth grew largely from material added by impacts from planetesimals, rather than by capture of dust grains from the solar nebula — or at least that's the inference from the latest geochemical analyses.

A record of Earth's formation and its evolutionary history during the past 4,500 million years is preserved within the chemical and isotopic composition of the mantle. Fluids and the magmas expelled at the Earth's surface as basalt rocks provide samples for deciphering this record. In particular, isotopes of the noble gases contain unique clues to the structure of the mantle, the formation of the hydrosphere and atmosphere, and the history of the building blocks used during our planet's accretion. On page 33 of this issue, Ballentine *et al.*¹ provide high-precision measurements of neon and helium isotopes in carbon-dioxide-rich well gases from New Mexico. Their results illuminate all of these issues, and have profound implications for our understanding of Earth's accretion history.

The initial (primordial) noble gases in the Earth were either trapped directly from a gas-rich solar nebula, or implanted as ions during intense irradiation by a young Sun². Terrestrial noble gases differ in their isotopic make-up from primordial values because they have been modified by the radioactive decay of uranium (U), thorium (Th) and potassium (K), the major heat-producing elements. The ratio of primordial to radiogenic noble gases in Earth's mantle therefore reflects the time-integrated ratio of primordial noble gas to U, Th and K. For example, the relatively high ratios of helium isotopes ($^3\text{He}/^4\text{He}$) observed in ocean island basalts (OIBs) from localities such as Hawaii and Iceland indicate a mantle source that is characterized by high $^3\text{He}/(\text{U}+\text{Th})$. This OIB source has a higher $^3\text{He}/^4\text{He}$ than that of mid-ocean-ridge basalts (MORBs), and is therefore less degassed and generally considered to lie somewhere below the upper mantle³.

Support for this model is found by comparing the neon-isotope compositions of OIBs and MORBs^{4–9}. Elevated $^{21}\text{Ne}/^{22}\text{Ne}$ is a result of ^{21}Ne production by nuclear processes involving the collision of energetic α -particles (^4He atoms produced by U and Th radioactive decay) with ^{18}O in mantle silicates — the silicon- and oxygen-rich rocks that make up most of the mantle. Hence, the trend in OIBs from Hawaii and Iceland^{6–8}, towards high $^{20}\text{Ne}/^{22}\text{Ne}$ and low $^{21}\text{Ne}/^{22}\text{Ne}$ when compared with MORBs^{4,5,9} (Fig. 1), is consistent with a deep, relatively unde-gassed 'mantle plume' source beneath those ocean islands. Elevated $^{20}\text{Ne}/^{22}\text{Ne}$ cannot be explained by nucleogenic processes, and

is attributed to the presence of a solar neon component in the Earth^{4–10}. A major goal is therefore to identify the upper limit for $^{20}\text{Ne}/^{22}\text{Ne}$ in various parts of the mantle, because this potentially distinguishes between different accretion scenarios for the Earth⁷.

Ballentine and colleagues' results¹ establish an upper limit of 12.2 to 12.5 for $^{20}\text{Ne}/^{22}\text{Ne}$ in Earth's upper mantle. In contrast, $^{20}\text{Ne}/^{22}\text{Ne}$ ratios for the deep mantle, estimated from analyses of basalts at Hawaii and Iceland^{6–8}, and rocks from the mantle-plume province of Russia's Kola Peninsula¹⁰, extend to 13.0 or higher. These higher $^{20}\text{Ne}/^{22}\text{Ne}$ values approach the value for the solar wind (13.8), a present-day proxy for the early solar nebula. The shallow- and deep-mantle sources are systematically different in

$^{21}\text{Ne}/^{22}\text{Ne}$ as well (upper mantle, 0.056; deep mantle, <0.04). The primordial neon-isotope composition for the upper mantle strongly resembles the neon component (Ne-B) observed in meteorites that underwent significant ion implantation by solar energetic particles (SEPs). Therefore, the primordial Ne-isotope composition of the deep mantle (OIB source) resembles that produced by direct trapping from a gas-rich solar nebula, whereas the primordial Ne-isotope composition of the upper mantle (MORB source) resembles that produced by a mixture of solar wind and SEPs (Fig. 1).

These measurements suggest that deep-mantle sources, such as those beneath Hawaii and Iceland, do not contribute much to the inventory of noble gases in the convecting upper mantle. Evidently, steady-state models for upper-mantle noble gases¹¹ that invoke a flux from these deep-mantle sources need to be re-evaluated.

More remarkably, however, the results indicate that accretion of the outer portions of the Earth was dominated by aggregated solids (planetesimals) that had been heavily irradiated by solar ions. This is remarkable because such intense irradiation is likely to have occurred during an active phase of the

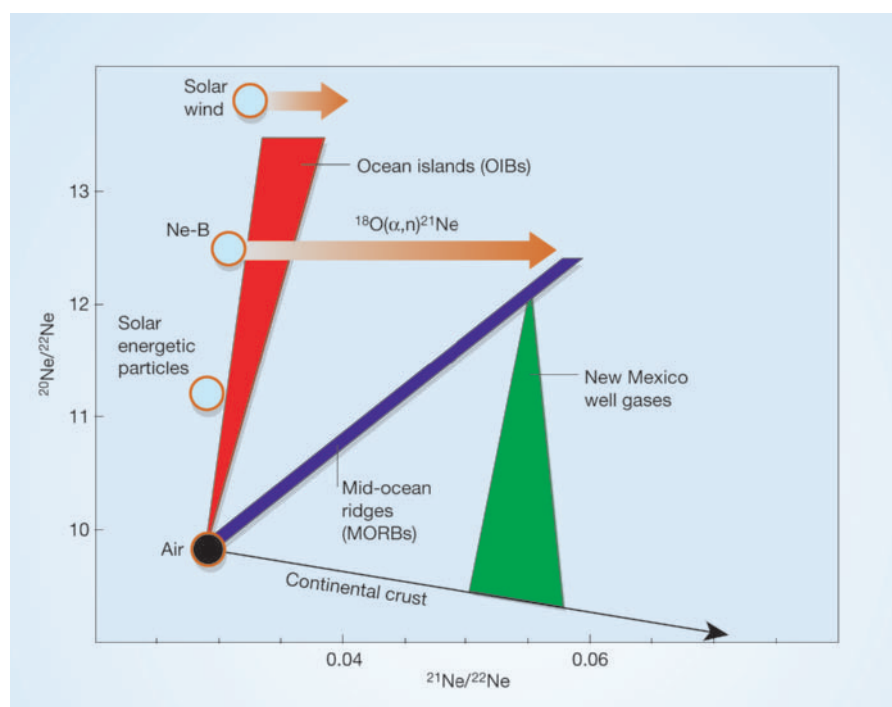


Figure 1 A three-component isotope mix. The diagram illustrates the neon-isotope compositions for air, solar energetic particles, Ne-B (the neon component in meteorites that underwent ion implantation by solar energetic particles) and the solar wind. Basalt rocks from mid-ocean ridges (MORBs) and ocean islands (OIBs) have air as one end-member because they contain atmospheric contamination released during mass spectrometric analysis. The OIBs from Hawaii and Iceland define a mixing line between air and a deep-mantle component similar to that of the solar wind, and MORBs define a mixing line between air and a primordial neon-isotope component that has been modified by addition of nucleogenic ^{21}Ne from α -particle collisions with ^{18}O in mantle silicates (arrows). Ballentine and colleagues' data¹ for well gases from New Mexico define a wedge-shaped field, because air- and crustal-derived neon are pre-mixed before the addition of gases from the upper mantle. This provides an estimated upper limit for $^{20}\text{Ne}/^{22}\text{Ne}$ in the upper mantle, which implicates Ne-B as the primordial composition for most of the mantle.

early Sun, and only after the rotating disk of nebular gas had been swept clear. The effects of this process have recently been imaged around the main-sequence star β Pictoris, where sub-micrometre dust has been swept out of this extrasolar planetary system by radiation pressure¹². With respect to Earth, only the deep-mantle regions feeding ocean islands such as Hawaii and Iceland seem to retain a considerable remnant of gases from the early solar nebula, captured from a dense atmosphere during the earliest parts of planetary formation.

One outstanding problem in this research is achieving a self-consistent model that incorporates the noble-gas constraints together with trace-element and isotope ratios of lithophile elements (those elements that tend to be concentrated in silicates, such as the alkaline earths and rare earths). The new neon-isotope results suggest that there is little or no exchange between the deep-mantle regions feeding ocean islands and the upper mantle. Yet there is currently no evidence in the lithophile tracers for any vestiges of primitive, undifferentiated mantle¹³. Evidence emerging from tungsten isotopes in oceanic basalts also seems to exclude significant interaction between the core and deep mantle¹⁴, making it unlikely that the core is the ultimate source of the solar neon-isotope signature observed in mantle plumes.

Consequently, the ultimate source seems to be remnants of the very earliest silicates involved in terrestrial accretion, and these remnants have remained effectively isolated from overlying mantle convection throughout Earth's history. If this source is associated with the seismically anomalous (D'') layer at the base of the mantle, the neon-isotope results indicate that this layer may have formed during Earth's accretion¹⁵.

David W. Graham is at the College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, Oregon 97331, USA.

e-mail: dgraham@coas.oregonstate.edu

- Ballentine, C. J., Marty, B., Sherwood Lollar, B. & Cassidy, M. *Nature* **433**, 33–38 (2005).
- Pepin, R. O. & Porcelli, D. *Rev. Mineral Geochem.* **47**, 191–246 (2002).
- Kurz, M. D., Jenkins, W. J. & Hart, S. R. *Nature* **297**, 43–47 (1982).
- Sarda, P., Staudacher, T. & Allègre, C. J. *Earth Planet. Sci. Lett.* **91**, 73–88 (1988).
- Marty, B. *Earth Planet. Sci. Lett.* **94**, 45–56 (1989).
- Honda, M., McDougall, I., Patterson, D. B., Douglis, A. & Clague, D. A. *Nature* **349**, 149–151 (1991).
- Trieloff, M., Kunz, J., Clague, D. A., Harrison, D. & Allègre, C. J. *Science* **288**, 1036–1038 (2000).
- Moreira, M., Breddam, K., Curtice, J. & Kurz, M. D. *Earth Planet. Sci. Lett.* **185**, 15–23 (2001).
- Moreira, M., Kunz, J. & Allègre, C. J. *Science* **279**, 1178–1181 (1998).
- Yokochi, R. & Marty, B. *Earth Planet. Sci. Lett.* **225**, 77–88 (2004).
- Porcelli, D. & Wasserburg, G. J. *Geochim. Cosmochim. Acta* **59**, 4921–4937 (1995).
- Okamoto, Y. K. *et al.* *Nature* **431**, 660–663 (2004).
- Hofmann, A. W. *Nature* **385**, 219–228 (1997).
- Scherstén, A., Elliott, T., Hawkesworth, C. & Norman, M. *Nature* **427**, 234–236 (2004).
- Tolstikhin, I. & Hofmann, A. W. *Phys. Earth Planet. Inter.* (in the press).

Human immunodeficiency virus

Nuclear RNA export unwound

Bryan R. Cullen

The ways in which HIV can subvert cellular processes for its own ends seem boundless. The latest discovery — a cellular enzyme that helps to export HIV RNA from the nucleus — reveals a possible drug target.

Gene expression is a multi-step process, the first stage of which is the production of a messenger RNA transcript of a gene. That mRNA is then used as a template to produce a protein. Here, cells whose genetic material is encased in a nucleus (eukaryotic cells) face a problem: their genes are transcribed in the nucleus but proteins are made outside it, in the cytoplasm, so the mRNA must be exported.

The same stricture applies to HIV, whose genes, once incorporated into the host genetic material, are likewise transcribed in the nucleus. It is known that the viral protein Rev recruits the cellular export factor Crm1 to export several essential HIV-1 mRNAs¹. Host mRNAs, by contrast, generally rely on another export factor, the Tap–Nxt1 dimer. But this is not the only requirement — a 'remodelling' enzyme is also needed². For cellular mRNAs, that enzyme is thought^{3–5} to be Dbp5, but this has no role in mRNA export mediated by Rev–Crm1. Writing in *Cell*, however, Yedavalli *et al.*⁶ suggest that the necessary enzyme is a member of the same family.

During transcription and processing, eukaryotic mRNAs are loaded with a wide variety of nuclear proteins that regulate the export, cytoplasmic localization, translation and stability of the mRNAs². Although some of these RNA-binding proteins work by remaining associated, at least transiently, with newly exported mRNAs, many others dissociate during or immediately after export. For most cellular mRNAs, this remodelling step is thought to be mediated, at least in part, by Dbp5, an RNA helicase belonging to the ubiquitous DEAD-box protein family^{3–5}.

DEAD-box helicases use the energy released from the hydrolysis of adenosine triphosphate (ATP) to unwind RNA structures and, perhaps more importantly in this context, to dissociate RNA–protein complexes^{7,8}. These enzymes each bear nine conserved amino-acid motifs, including the eponymous DEAD box itself — named after the abbreviations for the amino acids aspartic acid, glutamic acid and alanine — and computer analysis has used these motifs to identify numerous family members in all

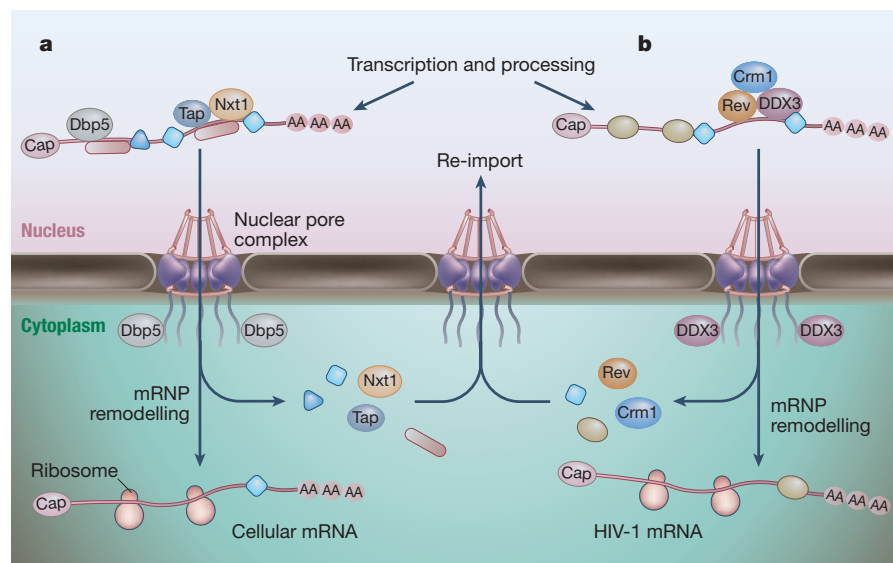


Figure 1 Escape from the nucleus: how cellular and HIV-1 messenger RNAs are exported.

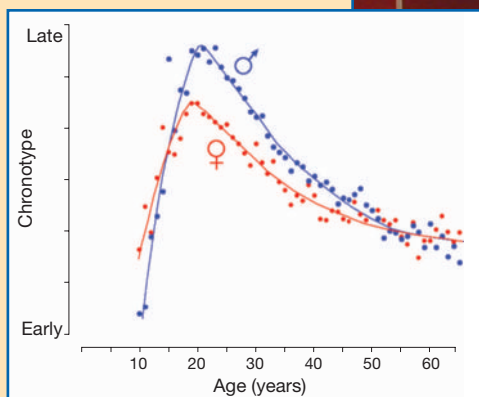
a, Transcription and processing of cellular mRNAs give rise to export-competent mRNA–protein (mRNP) complexes that include a range of proteins, including the export factors Tap and Nxt1 and an RNA helicase, Dbp5. During transport through nuclear pore complexes (NPCs), the mRNP is remodelled by pre-bound and/or NPC-associated Dbp5, releasing a range of factors that then return to the nucleus. The string of A's represents the polyadenosine tail characteristic of mRNAs. b, HIV-1 mRNAs are bound by a distinct but overlapping set of nuclear proteins that includes the export factors Rev and Crm1 and, as the findings of Yedavalli *et al.* now suggest⁶, the helicase DDX3. This mRNP is also remodelled during export, in this case by pre-bound and/or NPC-associated DDX3. Ribosomes then translate the cellular and HIV-1 mRNAs into protein.

Physiology

An end to adolescence

'Puberty' and 'adolescence' are not synonyms, although both terms describe that awkward age between childhood and adulthood. Puberty is defined as the period during which the reproductive system matures. It has a clearly defined marker for when it ends: when bone growth ceases. Adolescence, by contrast, is part physiological, part psychological, part social construct. Chronobiologists joke that people suffer adolescence twice — once themselves, and again when their own children hit the teenage years. But, frustratingly, they have not been able to define precisely when it ends.

Till Roenneberg *et al.* now provide a suitable marker (*Curr. Biol.* **14**, R1038–R1039; 2004). Their questionnaire-based survey of the sleeping habits of 25,000 people between the ages of 10 and 90 confirms the normal distribution of chronotypes in the



population: 'owls', who go to sleep late and wake up late; 'larks', who go to bed early and wake early; and those who fall between the two (the majority of people).

The authors' survey also shows a remarkably robust relationship of this chronotypology to age (see picture). Children are typically early risers but start to sleep

progressively later as they enter adolescence. They reach maximum lateness at around the age of 20, when the curve abruptly starts to decline. The normal distribution of owl-ness and lark-ness in the population is apparent at each age — but, the authors suggest, it is the point at which an individual's curve inverses that marks the end of adolescence.

The data also reflect the earlier development of females compared with males; young women reach their maximum lateness at 19.5 years and young men at 20.9 years. Females tend to be earlier chronotypes than men throughout adulthood, but the gender difference disappears at around 50 — the onset of menopause.

Alison Abbott

genomes analysed. The proteins have been proposed to participate in almost every process that involves RNA, including the splicing of precursor mRNAs, the production of the protein-synthesis machinery (ribosomes), mRNA translation, and of course nuclear mRNA export⁷.

Dbp5 was first genetically identified in yeast as a primarily cytoplasmic protein, enriched around the nuclear rim, that is essential for the export of cellular mRNAs from the nucleus^{3,4}. Subsequent analysis⁵ of the human protein revealed a specific association with the cytoplasmic side of nuclear pore complexes (NPCs), the large protein assemblies through which molecules go in and out of the nucleus.

More recently, evidence obtained in insect cells has revealed that Dbp5 actually shuttles between the nucleus and cytoplasm and, at least in part, assembles onto mRNA–protein complexes (mRNPs) as the mRNAs are being synthesized⁹. Importantly, microinjection of frog eggs with a mutant form of Dbp5 that could no longer hydrolyse ATP blocked the export of cellular mRNAs but did not affect mRNA export mediated by HIV-1 Rev and its cellular cofactor Crm1 (ref. 5). Therefore, whereas export of cellular mRNAs by the canonical Tap–Nxt1 pathway depends on Dbp5, the export of very similar viral mRNA substrates through the Rev–Crm1 pathway does not.

This finding suggested two possibilities:

either nuclear export mediated by Rev–Crm1 doesn't require mRNP remodelling, or it depends on a different RNA helicase. The second hypothesis has now been validated by Yedavalli and colleagues⁶. These authors show that overexpressing another member of the DEAD-box family, DDX3, in cultured human cells specifically enhances Rev–Crm1-dependent, but not Tap–Nxt1-dependent, nuclear mRNA export. Inhibiting DDX3 function selectively impairs the former process. Moreover, DDX3 interacts specifically with both Crm1 and Rev *in vivo*. Interestingly, DDX3, like Dbp5, is a nucleocytoplasmic shuttle protein that, at steady state, is predominantly localized at the cytoplasmic face of NPCs. In total, the data indicate that these two helicases, although specific for distinct RNA-export pathways, probably have a similar role in remodelling mRNPs during export (Fig. 1).

So what does DDX3, a cellular enzyme, do in uninfected cells? That remains unclear at present. Although Crm1 is essential for the export of HIV-1 mRNAs, it is not known whether it participates in the export of any cellular mRNAs. It does, however, have a well-established role in exporting proteins — as well as several non-protein-coding RNAs, including ribosomal RNAs¹ — from the nucleus. Yedavalli *et al.* present limited data suggesting that DDX3 is not required for Crm1-mediated protein export. But it might be involved in exporting certain

cellular coding or non-coding RNAs.

This issue is important, because Yedavalli *et al.*⁶ also present convincing evidence that DDX3 has an essential, and possibly rate-limiting, role in the HIV-1 life cycle. Thus they suggest that it might represent a new target for chemotherapeutic intervention. However, as cells undoubtedly did not evolve DDX3 solely for the convenience of this virus, it remains possible that inhibiting it could have unforeseen and harmful consequences. It would therefore seem to be a priority to knock down DDX3 expression in infected and uninfected T cells, the major cell type affected by HIV-1, as an initial test of the hypothesis that this helicase is indeed a worthwhile target for antiviral drug development.

Bryan R. Cullen is in the Department of Molecular Genetics and Microbiology, and the Center for Virology, Duke University Medical Center, Durham, North Carolina 27710, USA.

e-mail: culle002@mc.duke.edu

1. Cullen, B. R. *J. Cell Sci.* **116**, 587–597 (2003).
2. Dreyfuss, G., Kim, V. N. & Kataoka, N. *Nature Rev. Mol. Cell Biol.* **3**, 195–205 (2002).
3. Snay-Hodge, C. A., Colot, H. V., Goldstein, A. L. & Cole, C. N. *EMBO J.* **17**, 2663–2676 (1998).
4. Tseng, S. S.-I. *et al.* *EMBO J.* **17**, 2651–2662 (1998).
5. Schmitt, C. *et al.* *EMBO J.* **18**, 4332–4347 (1999).
6. Yedavalli, V. S. R. K., Neuveut, C., Chi, Y.-h., Kleiman, L. & Jeang, K.-T. *Cell* **119**, 381–392 (2004).
7. Rocak, S. & Linder, P. *Nature Rev. Mol. Cell Biol.* **5**, 232–241 (2004).
8. Schwer, B. *Nature Struct. Biol.* **8**, 113–116 (2001).
9. Zhao, J., Jin, S.-B., Björkroth, B., Wieslander, L. & Daneholt, B. *EMBO J.* **21**, 1177–1187 (2002).

Obituary

John R. Vane (1927–2004)

Sir John Robert Vane, who died on 19 November 2004, was one of the most distinguished British pharmacologists of all time. His greatest scientific achievement was the invention of the blood-bathed superfusion bioassay technique and its application to a number of seminal discoveries.

Vane used to enjoy saying that “bioassay measures biological activity” — an apparently tautological statement that hides the tremendous potential of the technique. For many years, pharmacological bioassay had been mainly carried out by immersing a piece of tissue in a physiological solution and studying its responses to biologically active agents. This method was taken further by B. Finkelman in the 1930s and J. H. Gaddum in the 1950s, who both developed the idea that a tissue could be superfused rather than bathed in a physiological solution, thereby increasing the sensitivity of the bioassay.

In the 1950s, when Vane was a young researcher at The Royal College of Surgeons, he modified the bioassay system so that different tissues could be suspended, one above the other, with the superfusion fluid flowing over each tissue in turn. By doing so, he developed a system whereby the differential sensitivity of three or four tissues gives a unique ‘fingerprint’ for each substance investigated. Most importantly, this system allowed the immediate detection of previously unrecognized biologically active agents. He also realized that, instead of a physiological solution, blood obtained directly from an animal could be used in an external circulation system for the instantaneous monitoring of the fate of bioactive materials.

In the early part of his career, Vane studied the way in which vasoactive substances — those that cause dilation or contraction of blood vessels — are handled by the circulation. This led him to propose that the lungs are metabolically active organs involved in controlling the concentration of certain local hormones or ‘autacoids’. This was a major contribution which spawned a great deal of research around the world.

By the early 1970s, Vane’s bioassay technique was regularly used in his laboratory to investigate the generation and fate of such agents as catecholamines, bradykinin, angiotensin and, most significantly, prostaglandins — hormone-like substances, the functions of which are now known to include modulation of



Guiding light in pharmacology

vascular tone, inflammation and pain. There was also the intriguing discovery of an ephemeral substance, released from the lungs in anaphylactic shock, which was recognized for its ability to contract rabbit aortic strips and was therefore given the acronym RCS (rabbit aorta-contracting substance). The release of this material was inhibited by aspirin-like drugs.

Soon after my arrival in 1971 at The Royal College of Surgeons, Vane invited me to join the project that led to the discovery that aspirin and aspirin-like drugs inhibit prostaglandin biosynthesis. The importance of this finding, undoubtedly Vane’s major contribution to biomedicine, cannot be overstated. Besides clarifying the mechanism of action of a drug widely used as an analgesic and anti-inflammatory agent for almost a hundred years, it pointed the way to treatments for various diseases, and opened avenues of research and drug discovery that are still being explored.

Notably, in 1974 RCS was identified as thromboxane A_2 , which is derived from the same precursors as prostaglandins: it is released from platelets in the blood and has powerful platelet-aggregating properties. This led to the unravelling of the mechanism of the now well-established anti-thrombotic properties of aspirin and to its successful use in the prevention and treatment of cardiovascular diseases. Vane used to muse that the 1969 experiments on RCS had already indicated aspirin’s mechanism of action,

because the responses of the tissues detecting prostaglandins were also reduced following aspirin treatment.

In 1973, Vane joined the Wellcome Foundation as director of research and development worldwide, and so took on responsibility for more than a thousand scientists. This distanced him from direct involvement in research, and he often complained to me privately about the burden of that job as compared with the delights of being close to the bench. Nevertheless, continuing in his footsteps, his group made further discoveries, significantly that of prostacyclin — a prostaglandin produced in the walls of blood vessels that acts as a vasodilator and inhibits platelet aggregation. This was a milestone in our understanding of vascular biology.

In 1985, Vane left Wellcome and returned to academic life. He then not only rebuilt a successful research team but established what is now a prestigious institute, the William Harvey Research Institute at St Bartholomew’s Hospital Medical School in London.

Although essentially a shy man, John Vane was charismatic and approachable, with a talent for creating an atmosphere in which intense work was combined with the most enjoyable social occasions. People from all over the world joined the group and were made to feel welcome — in many cases this led to lifelong friendships. During the years we worked together, some of our most productive discussions and ideas arose at the home of John and his wife Daphne in conversations after lunch or before dinner, or when travelling abroad to meetings.

John was an ingenious, hands-on pharmacologist, able to generate meaningful hypotheses almost effortlessly. Although his early training was in chemistry, he developed a great understanding of biological processes and a keen eye for the behaviour of the tissues in his beloved bioassay. One of his favourite phrases to students describing results they did not understand was “the tissues never lie” — it is the interpretation that can fail. Together with Bengt Samuelsson and Sune Bergström, he was awarded the Nobel Prize in Physiology or Medicine in 1982, and received many other accolades. Science in general, and the British Pharmacological Society and his friends and colleagues in particular, have lost a guiding light.

Salvador Moncada

Salvador Moncada is at the Wolfson Institute for Biomedical Research, University College London, Gower Street, London WC1E 6BT, UK. e-mail: s.moncada@ucl.ac.uk

Biomedical materials

Networks for artificial tissues

Adv. Mater. **16**, 2007–2012 (2004)

Porous, biodegradable polymers can provide scaffolding for tissue engineering: seeded with cells, they are broken down as the cell colony grows and takes on the shape of the scaffold. But the results are far from ideal. Real tissues typically have precise spatial arrangements of different cell types, supported by a vascular blood supply.

Kevin R. King *et al.* offer an improved approach by using three-dimensional microfluidic networks made from biodegradable polymers to deliver cells to precise locations, or to provide a built-in system of vascular channels. They imprint thin films of the polymer poly(lactic-co-glycolide) (PLGA) with grooves as narrow as 2 μm across, and then laminate the films into blocks laced with a precise network of channels. An innovation is in patterning the PLGA from a melt, rather than from solution, after which the soft material is pressed into a preformed rubbery mould. King *et al.* also use heat to weld successive layers together, creating robust joins with little thermal degradation of the polymer.

Such microstructured materials might be additionally useful for delivering drugs or growth factors in an implanted medical device or scaffold.

Phillip Ball

Chemical biology

Sweetness on a chip

Chem. Biol. **11**, 1701–1707 (2004)

Cell-surface carbohydrates are involved in many biological processes, including cell–cell recognition, and cell adhesion and signalling. These carbohydrates are structurally complex, making it difficult to study some of the processes involved at the molecular level. But many new tools are being developed for studying glycobiology, including carbohydrate microarrays. Matthew D. Disney and Peter H. Seeberger report that these biochips can be used to examine the interactions of live bacteria with carbohydrates — often a key step in the bacterial invasion of host cells.

The authors created slides containing various carbohydrates, and then treated these microarrays with *Escherichia coli* bacteria that had been labelled with a dye. They found that the bacteria bound strongly to the sugar mannose but not to other carbohydrates, and that few bacterial cells were needed to produce a strong signal. The non-destructive nature of the microarrays also allowed the authors to collect the bacteria, grow them



Flu shot. But ageing helper T cells render vaccines less effective in the elderly than the young.

in culture, and determine their susceptibility to various antibiotics. Disney and Seeberger went on to use the microarrays to screen for inhibitors of carbohydrate–cell interactions. They envisage that the biochips will be useful for studying pathogenic bacteria and developing new drugs to combat these microbes.

Joshua Finkelstein

Astronomy

Focus on exoplanets

Astrophys. J. Lett. **618**, L165–L168 (2005)

In the hunt for planets outside our Solar System, visual identification is the most difficult method to apply, but potentially the most informative. The problem is that at visible wavelengths an exoplanet image will be typically about a billion times fainter than its parent star, and a million times less bright in the infrared. The star's light also produces diffraction rings in conventional telescopes that often extend far beyond the orbit of an associated planet.

A. H. Greenaway *et al.* outline an ancillary optical system that could help to surmount these difficulties. Their 'pupil replication' method reduces the brightness of the star's diffraction rings, making it easier to cut out the starlight that swamps a planet's image.

Computer simulations show that the system can reduce by a factor of three the angle between a planet and star in which the planet can be detected. It should also be much cheaper and easier to implement than one of the alternative options — building a telescope with a larger diameter.

Mark Peplow

Immunology

Helpless B cells

J. Exp. Med. **200**, 1613–1622 (2004)

Elderly people are strongly encouraged to have vaccinations against infectious diseases such as influenza and pneumonia. But their bodies are not always up to the task; their immune systems produce relatively weak and short-lived antibody responses, which make the shots less effective. Sheri M. Eaton *et al.* have investigated whether this defective reaction involves an age-related problem with the antibody-producing B cells, or with the T cells that help them out.

When the authors transferred aged helper T cells to young mice, the animals' B-cell response to immunization was crippled. But young and old mice given helper T cells from young donors showed no difference in B-cell activity. The findings suggest that helper T cells are the weakest link in older immune systems. According to Eaton *et al.*, a better understanding of these age-related defects could help to improve vaccine efficacy in the elderly.

Roxanne Khamis

Molecular biology

Silence, please

Cell **119**, 941–953 (2004)

Every cell nucleus is crammed with about two metres of DNA, wound snugly around proteins called histones to form chromatin. Just how tightly the chromatin is coiled in a particular region determines the availability of its genes; active genes are usually in looser, more open sections. Yujiang Shi *et al.* now add a new variety to the list of enzymes that control how open the chromatin is.

Chemical modifications to the tails of histone proteins are crucial to chromatin structure. Several types of modification change continually, with the balance between modified and unmodified histones controlling chromatin structure. Modification of histones by methylation has been considered more permanent, because only enzymes that add methyl groups (methylases) were known. But Shi *et al.* have discovered a histone demethylase.

Called LSD1 (for lysine-specific demethylase 1), this enzyme removes methyl groups from a specific lysine amino acid in histone protein H3. Methylation of this residue has been associated with active genes, so it is unsurprising that LSD1 activity shuts down certain genes. Moreover, Shi *et al.* find that decreasing the amount of LSD1 allows the genes to be turned on again. It seems that histone methylation is a dynamic process after all, and depends on the balance between methylases and demethylases.

Helen Dell

Neural processing of a whistled language

A rare surrogate of Spanish highlights the adaptability of the brain's language regions.

Silbo Gomero is a whistled language that is a rare and endangered surrogate of Spanish, used by shepherds on the island of La Gomera in the Canary Islands for communication over long distances on difficult terrain. Here we show that areas of the brain normally associated with spoken-language function are also activated in proficient whistlers, but not in controls, when they are listening to *Silbo Gomero*. Our findings demonstrate that the language-processing regions of the human brain can adapt to a surprisingly wide range of signalling forms.

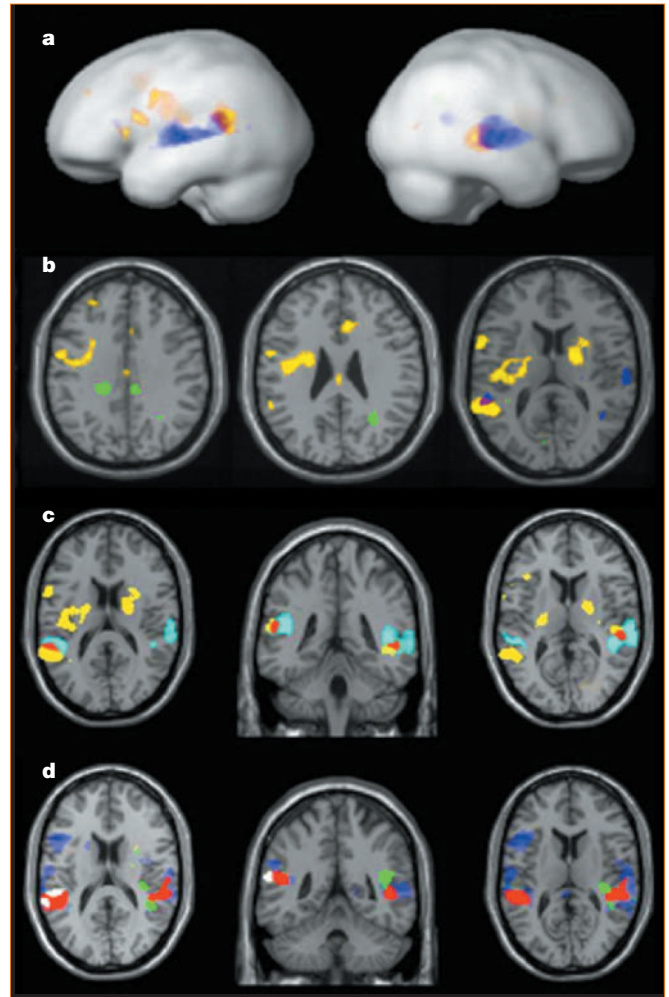
The traditionally recognized 'language' areas in the left temporal and inferior frontal lobes of the brain are not exclusive to speech processing — for example, they are engaged in the perception of visual-gestural linguistic signals in sign languages used by the deaf¹ and in non-linguistic acoustic signals^{2,3}. We have investigated which areas of the brain are activated in users of *Silbo Gomero*, henceforth *Silbo*, during their comprehension of this whistled language.

Silbo reduces the full phonemic inventory of Spanish to two phonologically contrasting vowels and four consonants⁴. Whistled 'words' are formed by recoding the vowels and consonants of individual Spanish words into whistles that vary along a pitch dimension (high to low) and which differ with respect to the character of the melodic line (continuous or interrupted). Although this leads to phonological mergers and hence potential ambiguities, in practice users rely on repetition and context for the communication of short and simple, routine messages. The compositional, formant-like glides of *Silbo* can therefore function as a form of linguistic communication, provided that listeners know the rules of the whistled codification and can interpret the semantic content in the shared cultural context.

We acquired functional neuroimaging data while users (*Silbadores*) and non-users (controls) of *Silbo* were exposed to comprehension tasks (for details, see supplementary information). The first task involved listening passively to *Silbo* and to Spanish sentences against a baseline condition of digitally reversed *Silbo*. In the second task, participants were asked to monitor cycles of *Silbo* 'words' and Spanish words intermixed with silent periods.

Our results show that the temporal regions of the left hemisphere that are usually associated with spoken-language function^{5–7} are engaged during the processing of *Silbo* in experienced *Silbadores* (Fig. 1). Passive-listening and active-monitoring tasks produce

Figure 1 Colour-coded brain-activation patterns from functional neuroimaging of *Silbadores* and non-whistler controls, produced in response to tasks in *Silbo Gomero* and in Spanish. **a**, Surface rendering of each side of the brain, and **b**, axial sections of a normalized brain, showing sites of activation by passive listening to sentences in *Silbo* and in Spanish. **c**, **d**, Axial and coronal (centre) sections showing sites of activation for **c**, *Silbadores* in *Silbo* passive-listening and *Silbo*-monitoring tasks; and for **d**, *Silbadores* and non-whistlers in all *Silbo* and all Spanish-spoken tasks. Voxel coloration for activation response: yellow, *Silbadores* to *Silbo* sentences (**a–c**); green, non-whistlers to *Silbo* sentences (**b**) and to all *Silbo* tasks (**d**); dark blue, all subjects to Spanish sentences (**a**, **b**) and to spoken Spanish tasks (**d**); light blue, *Silbadores* to *Silbo*-monitoring tasks; white, *Silbadores* to all *Silbo* tasks; and red, common activations in *Silbadores* for *Silbo* and Spanish (**a**, **b**, **d**) and for the two *Silbo* tasks in **c**. The left side of the axial and coronal images corresponds to the left side of brain.



a common activation in the left superior posterior temporal gyrus, near the temporal-parietal junction (Fig. 1c); activation of the right-hemisphere superior-midtemporal region is also evident across both the *Silbo* and Spanish speech conditions (Fig. 1a, d).

Activity increases in the right temporal lobe in response to non-linguistic pitch changes, tones and complex sounds^{8,9}, but the same regions may also be associated with linguistic processing tasks — particularly at the sentence level¹⁰. However, we identified no common cortical language areas for *Silbo* and for speech in non-whistlers (Fig. 1a, b). Group analysis indicated that the areas activated during both Spanish and *Silbo* processing in *Silbadores* differ significantly from those in non-whistlers (Fig. 1d, and see supplementary information). A time-series analysis of the region of interest for each subject verified that *Silbo* modulates cortical activity only in the *Silbadores* and not in the controls (see supplementary information).

How is this pattern of activation for whistle processing in experienced users of *Silbo* (Fig. 2) explained? On the one hand, whistled speech relies upon changes in pitch and melodic form to create distinctive acoustic patterns; on the other, it serves a communicative function. Our results indicate that, in this situation, it is the temporal-lobe regions implicated in language processing that respond, even though the signal is a whistle and the language an unusual speech surrogate.

Left-hemisphere temporal and parietal regions may provide complementary pathways for language processing^{5–7,11}. It has been proposed that an anterior-ventral system is used for analysing and mapping complex acoustic sounds, such as speech, onto lexical representations and that a posterior-dorsal system processes the articulatory-gestural representation of speech⁸. We saw less ventral-anterior temporal activation during *Silbo*-processing than during speech-processing tasks, perhaps because this simple whistled system has only a limited number

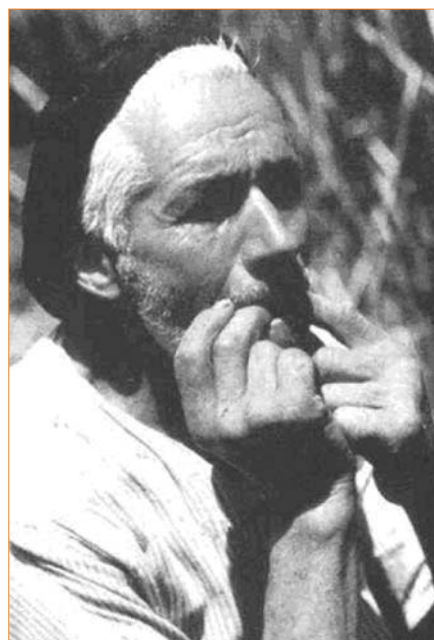


Figure 2 Silbador from La Gomera in the Canary Islands using the whistled language *Silbo Gomero* as a means of remote communication. The language recodes the vowels and consonants of individual Spanish words into whistles.

of uniquely specified phonological contrasts that need to be analysed. But for both speech and *Silbo* processing, the posterior temporal cortex is activated in a region that is involved in articulatory–gestural representations^{12,13}. The presence of premotor activation (tongue and lip representation) that is involved during *Silbo* communication is consistent with this interpretation.

Manuel Carreiras*, Jorge Lopez†, Francisco Rivero*, David Corina‡

*Departamento de Psicología Cognitiva, Universidad de La Laguna, La Laguna, Tenerife, 38205, Spain

e-mail: mcarreir@ull.es

†IMETISA, Hospital Universitario de Canarias, La Laguna, Tenerife, 28320, Spain

‡Department of Psychology, University of Washington, Seattle, Washington 98195–1525, USA

- Petito, L. A. et al. *Proc. Natl Acad. Sci. USA* **97**, 13961–13966 (2000).
- Joanisse, M. F. & Gati, J. S. *NeuroImage* **19**, 64–79 (2003).
- Giraud, A. L. & Price, C. J. *J. Cogn. Neurosci.* **13**, 754–765 (2001).
- Trujillo, R. *El Silbo Gomero: Análisis Lingüístico* (Canaria-Andrés Bello, Tenerife, 1978).
- Hickok, G. & Poeppel, D. *Trends Cogn. Sci.* **4**, 131–138 (2000).
- Binder, J. R. et al. *Cereb. Cortex* **10**, 512–528 (2000).
- Scott, S. K. & Johnsrude, I. S. *Trends Neurosci.* **26**, 100–107 (2003).
- Specht, K. & Reul, J. *NeuroImage* **20**, 1944–1954 (2003).
- Warren, J. D. & Griffiths, T. D. *J. Neurosci.* **23**, 5799–5804 (2003).
- Kircher, T. T. J., Brammer, M., Andreu, N. T., Williams, S. C. R. & McGuire, P. K. *Neuropsychologia* **39**, 798–809 (2001).
- Belin, P. & Zatorre, R. J. *Nature Neurosci.* **3**, 965–966 (2000).
- Hickok, G., Buchsbaum, B., Humphries, C. & Muftuler, T. *J. Cogn. Neurosci.* **15**, 673–682 (2001).
- Hickok, G. & Poeppel, D. *Cognition* **92**, 67–99 (2004).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Environment

Early ant plagues in the New World

The identity and origin of the West Indian plague ants of the early sixteenth and late eighteenth centuries have long been a mystery¹. By reviewing historic accounts with an analysis of the present-day Caribbean ant fauna, I have narrowed the list of suspects to two species and their insect symbionts.

During, or soon after, 1518–19, a plague of stinging ants hit the fledgling Spanish settlements on Hispaniola, the second largest island in the Greater Antilles. According to an eyewitness account by the premier colonial historian Bartolomé de Las Casas², the insects destroyed crops over a substantial portion of the island and invaded dwellings. During 1760–70, similar ant plagues spread through the Lesser Antilles, reducing sugarcane fields to “a state of the most deplorable condition”³.

What were the plague ants, and why did they multiply to such proportions? I investigated these questions during field work on the historic plague islands, paying particular attention to species that might be descendants of the plague ants. I also combed contemporary accounts of the ants and fitted together information concerning their appearance and habits, which I matched against that of the now reasonably well-known modern ant fauna. By a process of elimination, I narrowed the list of fauna to candidates with all of the available defining traits.

The Hispaniolan plague ant is easily characterized from the first-hand account of Las Casas². The ant he described was very aggressive; it had a painful sting; it occurred in dense populations in the root systems of shrubs and trees; it did not cut above-ground vegetation yet somehow damaged the root systems; and it was also a pest in houses and gardens. The only species also present in the modern West Indian ant fauna that has all these qualities is the tropical fire ant, *Solenopsis geminata*. This inference is strengthened by the fact that another fire-ant species, *S. invicta*, attained plague proportions in the Gulf states in the 1940s, following its introduction from the La Platte region of Brazil and northern Argentina⁴.

The 1760–70 plague ants of the Lesser Antilles had the same traits as *S. geminata* save two: in the several accounts of these species, including details from an eye-witness on Grenada⁵, there is not a single mention of defensive aggression or of stinging by the ants. An attack by swarms of fire ants is unavoidable if an intruder nears their nests, and would surely have been mentioned by anyone who had experienced it. And a fire-ant sting contains a venom that burns, creating a small, itching welt, which would

presumably also have been reported.

There is one possible clue: on Barbados in the mid-1600s there was an ant that was a serious house pest. Unlike fire ants, its workers lifted and carried large food items, such as cockroaches, in an unusually coordinated fashion. This feature points to the ant genus *Pheidole*. Among the many species known from the West Indies⁶, only two are candidates: *P. jelskii*, a native species, and *P. megacephala*, which is of African origin. The evidence favours *P. megacephala*, a global, invasive ant that has caused similar problems in other tropical countries⁷.

A puzzle remains: only the attine leafcutter ants of the New World are known to attack vegetation as an important source of food, yet entire plantations on Hispaniola were wiped out “as though fire had fallen from the sky and scorched them”, records Las Casas². Sugar cane in the fields of the Lesser Antilles likewise disappeared in the 1760s (ref. 3).

The only viable hypothesis is that the ants had a symbiotic relationship with insects that attack plants directly. The two plague-ant suspects, *S. geminata* and *P. megacephala*, heavily attend sap-sucking coccids, mealy bugs and other insects of the Homoptera group⁷. The ants protect these insects in exchange for their abundant excrement, which is rich in sugar and amino acids. The Spanish, not recognizing the role of the homopterous sap-suckers in the midst of the myriad kinds of insect teeming around their crops, would understandably put the blame on the stinging ants. It was not until the late eighteenth century, on Grenada, that naturalists began to suspect the involvement of homopterans in the West Indian ant plagues⁵.

Edward O. Wilson

Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, Massachusetts 02138-2902, USA

e-mail: ewilson@oeb.harvard.edu

- Wheeler, W. M. in *Translation of Histoire des Fourmis* (de Réaumur, R. A. F., unpublished, 1742–43) 232–239 (Knopf, New York, 1926).
- de Las Casas, B. in *Historia de las Indias III, Capítulo CXXXVIII*, 270–271 (Imp. de M. Ginesta, Madrid, 1875–1876).
- Schomburgk, R. H. in *History of Barbados* 640–643 (Longman, Brown, Green and Longmans, London, 1848).
- Lofgren, C. S. & Vander Meer, R. K. (eds) *Fire Ants and Leaf-cutting Ants: Biology and Management* (Westview, Boulder, Colorado, 1986).
- Castles, J. *Phil. Trans. R. Soc. Lond.* **16**, 688–694 (1804).
- Wilson, E. O. *Pheidole in the New World: A Dominant, Hyperdiverse Ant Genus* (Harvard Univ. Press, Cambridge, Massachusetts, 2003).
- Hölldobler, B. & Wilson, E. O. *The Ants* (Belknap, Harvard Univ. Press, Cambridge, Massachusetts, 1990).

Competing financial interests: declared none.

brieF communications arising online

♦ www.nature.com/bca

Human behaviour: Egalitarian motive and altruistic punishment

J. H. Fowler, T. Johnson, O. Smirnov

(doi:10.1038/nature03256)

Reply: E. Fehr, S. Gächter (doi:10.1038/nature03257)

Human behaviour

Egalitarian motive and altruistic punishment

Arising from: E. Fehr & S. Gächter *Nature* **415**, 137–140 (2002)

Altruistic punishment is a behaviour in which individuals punish others at a cost to themselves in order to provide a public good. Fehr and Gächter¹ present experimental evidence in humans indicating that negative emotions towards non-cooperators motivate punishment, which, in turn, provokes a high degree of cooperation. Using Fehr and Gächter's original data, we provide an alternative analysis of their experiment that suggests that egalitarian motives are more important than motives for punishing non-cooperative behaviour. This finding is consistent with evidence that humans may have an evolutionary incentive to punish the highest earners in order to promote equality, rather than cooperation².

In the experiment by Fehr and Gächter, groups with four members played a public-good game. Each participant was given an initial endowment of 20 money units (MUs), which they could either keep or contribute (entirely or partially) to a group project. For every MU invested in the project, each member earned 0.4 MU. Although the dominant strategy in the game is to keep the whole endowment, mutual contribution yields the best result for the group. In one treatment, subjects had an option to decrease the payoff of other group members, such that 1 MU spent on punishment decreased the payoff of the targeted individual by 3 MUs. The punishment stage started immediately after subjects had seen the payoffs earned by other group members in the first stage.

Punishment in the experiment was frequent and followed a pattern. Most negative points were imposed on below-average contributors and those who earned above-average payoffs in the first round. Fehr and Gächter define defection in relative terms, asserting that subjects punish an individual j in proportion to his or her deviance from the mean contribution of the other three players:

$$\frac{1}{3} \sum_{i \neq j} (c_i) - c_j$$

However, suppose individuals were not concerned about contributions and instead wanted to minimize inequality in the payoffs. If so, they might choose punishments in proportion to payoff deviance:

$$\pi_j - \frac{1}{3} \sum_i \pi_i$$

Notice that, as

$$\pi_j = 0.4 \sum_i (c_i) - c_j$$

in the Fehr and Gächter experiment, payoff

deviance is exactly equal to contribution deviance:

$$0.4 \sum_i (c_i) - c_j - \frac{1}{3} \sum_{i \neq j} \left(0.4 \sum_i (c_i) - c_j \right) = \frac{1}{3} \sum_{i \neq j} (c_i) - c_j$$

Thus, it is not possible to tell them apart, and all of Fehr and Gächter's statistical results equally support the hypothesis that subjects are punishing the top earners in order to minimize the difference in payoff outcomes.

If absolute levels are used instead of deviance from the mean, the experiment suggests that payoffs are important in altruistic punishment. We replicated Fehr and Gächter's regression analysis of the data and then used the same method to examine how group expenditures for the punishment of player i varied with player i 's contribution, prepunishment payoff, and an interaction between the two.

The resulting model suggests that the payoff has a strong and significant effect on punishment, even controlling for the contribution. For example, a 10-MU increase in the payoff yields 6.1 MU (± 1.1 MU) of additional punishment when the contribution is 0, and 1.8 MU (± 1.4 MU) when the contribution is 20. By contrast, the contribution has less effect on punishment and only decreases punishment when the payoff is sufficiently high. A 10-MU increase in the contribution yields a 3.6-MU (± 1.4 MU) decrease in the total punishment when the payoff is 44 (the maximum observed value), but the contribution has no significant effect on punishment when the payoff is 13 MU (the minimum observed value). These results indicate that subjects were more motivated to punish high earners than low contributors, and that egalitarian motives may underlie altruistic punishment in humans.

James H. Fowler*, **Tim Johnson†**, **Oleg Smirnov‡**

*Department of Political Science, University of California, Davis, One Shields Avenue, Davis, California 95616, USA
e-mail: jhfowler@ucdavis.edu

†Max Planck Institute for Human Development, Center for Adaptive Behaviour and Cognition, 14195 Berlin, Germany

‡Department of Political Science, University of Oregon, 1284 University of Oregon, Eugene, Oregon 97403, USA

doi:10.1038/nature03256

1. Fehr, E. & Gächter, S. *Nature* **415**, 137–140 (2002).

2. Boehm, C. *Hierarchy in the Forest: The Evolution of Egalitarian Behaviour* (Harvard Univ. Press, Cambridge, 1999).

Fehr and Gächter reply — Fowler *et al.* raise an important question¹. They correctly argue that the desire to reduce inequality may motivate cooperators who altruistically punish free riders in our experiments². Also, the evolutionary history of humans suggests that egalitarianism shaped many human cultures³ and that egalitarian motives may, therefore, be a powerful force behind the punishment of free riders. In addition, recently developed proximate theories⁴, which formalize the notion of inequality aversion, also suggest that egalitarian desires may be important. Fowler *et al.* contrast their egalitarianism hypothesis with our view that negative emotions against free riders drive punishment.

However, the two views are not necessarily incompatible: egalitarian sentiments may be the basis behind cooperators' negative emotions because free riding causes considerable inequalities. Moreover, the reanalysis of our original data by Fowler *et al.* can only raise (but not settle) the question of whether equality motives are important because a punishing cooperator in our experiments² inevitably reduces the inequality between himself and the punished free rider. Thus, it is not possible to isolate any other motive behind altruistic punishment based on these data because the equality motive can never be ruled out.

A plausible alternative to the egalitarian motive is that cooperative subjects may perceive free riding as a violation of the strong reciprocity norm^{5–7}. Cooperators may feel exploited by the free riders because the latter did not reciprocate their cooperative choices. Retaliation motives drive altruistic punishment in this view.

The retaliation motive has been isolated in a public-good experiment (A. Falk, E. F. and U. Fischbacher, see www.iew.unizh.ch/wp/iewwp059.pdf) in which the potential impact of the equality motive was removed. This experiment was almost identical to our original², except that punishment did not change the income difference between the punished and the punishing subject. One money unit (MU) spent on punishment reduced the free rider's payoff by exactly this amount. Thus, if egalitarian motives are the sole driving force behind altruistic punishment, there should be no punishment in this experiment. However, punishment is frequently observed (Fig. 1).

This punishment pattern is very similar to that of the original experiment because those who cooperate predominantly punish the free riders. Overall, subjects punish other group members in the new experiments 211 times: 51 out of 87 subjects (59%) punish at least once, and 22% punish more than five times during the experiment, which consists of six rounds. There is a considerable amount of punishment in the new experiments, although the equality motive cannot be

operating. However, the level of punishment in this experiment is considerably lower than in our original experiments². This is consistent with, but does not demonstrate, the importance of egalitarian motives, because the new experiment excludes egalitarian motives by making punishment more expensive. Therefore, the difference between the two experiments could be due to the higher cost of punishment or to the absence of egalitarian reasons for punishing.

At a minimum, these experiments indicate that the equality motive is not the sole driving force behind altruistic punishment. The retaliation motive also seems to be important, maybe even more so than the equality motive. Evidence from other cooperation⁸ and bargaining⁹ experiments supports this conclusion. In bargaining experiments, for example, subjects are much more likely to reject an unfair

offer made by a human subject than an identical offer generated by a random mechanism¹⁰. Likewise, subjects reject identical unequal human offers much

more often if more equal offers have been available, than they do if only more unequal offers have been available¹¹. Thus, the same level of inequality triggers very different rejection behaviours, depending on the source of the inequality and the set of available offers.

Ernst Fehr*, **Simon Gächter†**

*University of Zürich, Institute for Empirical Research in Economics, 8006 Zürich, Switzerland
e-mail: efehr@iew.unizh.ch

†University of St Gallen, FEW-HSG, 9000 St Gallen, Switzerland

doi:10.1038/nature03257

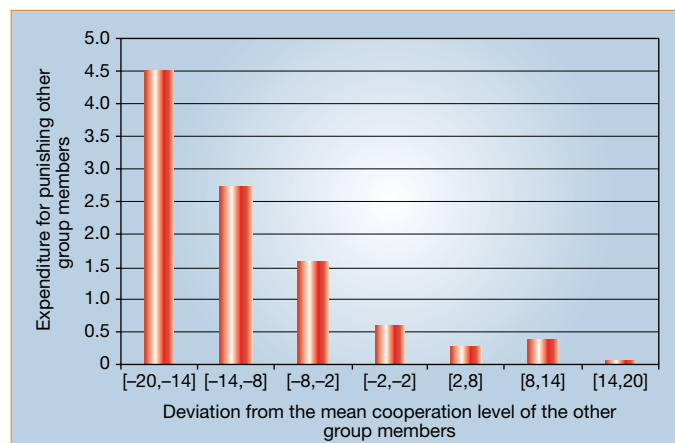


Figure 1 Mean expenditure on punishment as a function of the deviation of the punished group member's cooperation from the average cooperation of the other members. Each money unit (MU) spent on punishment reduced the punished member's income by 1 MU, so that punishment did not change the income difference between a punishing group member and a punished member. Therefore, the equality motive could not affect punishment in these experiments. The figure shows that altruistic punishment prevails even when punishment does not remedy inequality. For example, group members spend 4.5 MUs on punishing individuals whose contribution to the public good deviated between -20 and -14 units from the group-average contribution.

1. Fowler, J. H., Johnson, T. & Smirnov, O. *Nature* doi:10.1038/nature03256 (2004).
2. Fehr, E. & Gächter, S. *Nature* **415**, 137–140 (2002).
3. Boehm, C. *Curr. Anthropol.* **34**, 227–254 (1993).
4. Fehr, E. & Schmidt, K. M. *Q. J. Econ.* **114**, 817–868 (1999).
5. Gintis, H. *J. Theoret. Biol.* **206**, 169–179 (2000).
6. Bowles, S. & Gintis, H. *Theoret. Pop. Biol.* **65**, 17–28 (2004).
7. Falk, A. & Fischbacher, U. *Games Econ. Behav.* (in the press).
8. Anderson, C. M. & Putterman, L. *Games Econ. Behav.* (in the press).
9. Offerman, T. *Eur. Econ. Rev.* **46**, 1423–1437 (2002).
10. Blount, S. *Org. Behav. Hum. Dec. Proc.* **63**, 131–144 (1995).
11. Falk, A., Fehr, E. & Fischbacher, U. *Econ. Inq.* **41**, 20–26 (2003).

Neon isotopes constrain convection and volatile origin in the Earth's mantle

Chris J. Ballentine¹, Bernard Marty^{2,3}, Barbara Sherwood Lollar⁴ & Martin Cassidy⁵

¹Department of Earth Sciences, University of Manchester, Oxford Road, Manchester M13 9LP, UK

²Centre de Recherches Pétrographiques et Géochimiques, 15 Rue Notre Dame des Pauvres, BP 20, and ³Ecole Nationale Supérieure de Géologie, Rue du Doyen Roubault, 54501 Vandœuvre lès Nancy Cedex, France

⁴Department of Geology, University of Toronto, 22 Russell Street, Toronto, Ontario M5S 3B1, Canada

⁵Department of Geosciences, University of Houston, Houston, Texas 77204-5503, USA

Identifying the origin of primordial volatiles in the Earth's mantle provides a critical test between models that advocate magma-ocean equilibration with an early massive solar-nebula atmosphere and those that require subduction of volatiles implanted in late accreting material. Here we show that neon isotopes in the convecting mantle, resolved in magmatic CO₂ well gases, are consistent with a volatile source related to solar corpuscular irradiation of accreting material. This contrasts with recent results that indicated a solar-nebula origin for neon in mantle plume material, which is thought to be sampling the deep mantle. Neon isotope heterogeneity in different mantle sources suggests that models in which the plume source supplies the convecting mantle with its volatile inventory require revision. Although higher than accepted noble gas concentrations in the convecting mantle may reduce the need for a deep mantle volatile flux, any such flux must be dominated by the neon (and helium) isotopic signature of late accreting material.

The difference between the noble gas isotopic compositions of convecting mantle and deep mantle, sampled by mid-ocean-ridge volcanism and ocean island volcanism, respectively, has been a cornerstone of the 'layered mantle' model that has dominated our conceptual understanding of the terrestrial mantle for the past 25 years^{1,2}. The difference in the ³He/⁴He ratio between the values for plume-source basalts and the more uniform (but mostly lower) ³He/⁴He from mid-ocean-ridge basalts (MORB) was explained by a steady-state transfer of material from a primitive, volatile-rich 'lower' mantle into an 'upper' mantle, separated by the phase change at 670 km depth³. Further support for a layered mantle included the K-derived ⁴⁰Ar mass balance between the atmosphere and solid Earth, which pointed to a hidden reservoir with a ⁴⁰Ar concentration significantly higher than that in the upper mantle⁴. Similarly, the imbalance between heat and helium fluxes from the Earth was consistent with a mantle boundary layer (670 km) capable of separating these co-products of U and Th decay⁵.

Geoid and dynamic topography, seismic tomographic imaging and fluid dynamical studies, taken together, show that chemical layering is not achieved by the 670-km phase change^{6–8}. The existence of still deeper convectively isolated volatile-rich layers or regions to provide the volatile flux to the convecting mantle has been advocated^{9,10}. The compositional density contrast proposed to stabilize these regions should be observed seismically. Significantly, however, it has not been imaged¹¹. Other recent conceptual models include a water-rich melt layer within the mantle that preserves a volatile-rich deep reservoir while allowing whole-mantle convection¹². These, and the original steady-state models, source all primitive volatiles in the convecting mantle from a deep, volatile-rich reservoir.

Mantle-derived samples contain ²⁰Ne/²²Ne ratios higher than the atmospheric value (9.8), and are interpreted as evidence for trapped solar neon in the mantle². Indeed, the ²⁰Ne/²²Ne ratio of the Sun and the solar nebula is thought to be >13.4, derived from analysis of solar wind trapped in the lunar regolith (13.4–13.8), solar wind in Al foil (13.7 ± 0.3) and observation of the solar corona (13.8 ± 0.7)^{13,14}. These values are in distinct contrast to a mixture of SEP (solar energetic particle) Ne and solar-wind Ne found in meteoritic material irradiated by solar atoms and ions¹⁵, called

Ne-B¹⁵. The mixture of these two components is found in relatively uniform proportions to give a Ne-B value of ²⁰Ne/²²Ne = 12.5 ± 0.04 (ref. 16). Work reported in refs 16 and 17 has highlighted the importance of identifying the source of the Ne isotopes in different mantle reservoirs, as this information provides a critical evaluation of the mechanisms proposed to incorporate volatiles into the silicate Earth and, shown here, the extent of interaction between different mantle reservoirs. To date, however, interpretation has been compromised by ubiquitous air contamination found in MORB and ocean island basalt (OIB) samples¹⁸. We show in this work that magmatic natural gases can be used to obtain an unambiguous Ne isotopic value for the convecting mantle that is consistent with an irradiated meteorite origin (Ne-B). When compared to the highest reliable values found in deep mantle plume material (>13.0 ± 0.2)¹⁹, which are closer to solar nebula values, our result rules out the possibility that this OIB volatile source provides the noble gases found in the convecting mantle.

Magmatic CO₂ in New Mexico and noble gas results

Since the first identification of magmatic ³He in continental fluids, there has been an increasing awareness that magmatic CO₂ can dominate some crustal fluid systems^{20,21}. The Bravo dome natural gas CO₂ field (Fig. 1) was discovered in 1916 in Harding County, New Mexico²² (it was originally known as the Bueyeros field). Today this field is producing from over 250 wells. The gas is 98.6–99.8% CO₂ with trace amounts of N₂, CH₄ and noble gases. Earlier noble gas isotopic studies of one well in the old Bueyeros section^{23–25} show the CO₂ to be mantle-derived²³. Here, 15 samples were collected from producing wells across the field. Between 10 and 30 cm³ STP of sample gas was analysed²⁶. The ⁴He, ²⁰Ne, ⁴⁰Ar and ⁸⁴Kr abundances, and ³He/⁴He, ^{20,21}Ne/²²Ne and ⁴⁰Ar/³⁶Ar isotope ratios, were determined for each sample (Table 1). Xe isotopes, ³⁸Ar/³⁶Ar and stable isotope results will be presented in future publications.

³He/⁴He shows a coherent variation from 0.76R_a to 4.23R_a (R_a is the atmospheric ³He/⁴He = 1.4 × 10^{−6}) across the field (Fig. 1). ²⁰Ne/²²Ne and ⁴⁰Ar/³⁶Ar show the same coherent spatial variation, and have some of the highest values measured in a free crustal fluid, up to 11.88 and 22,600, respectively. The samples define a plane in three-dimensional plots of I/²²Ne versus ²¹Ne/²²Ne versus

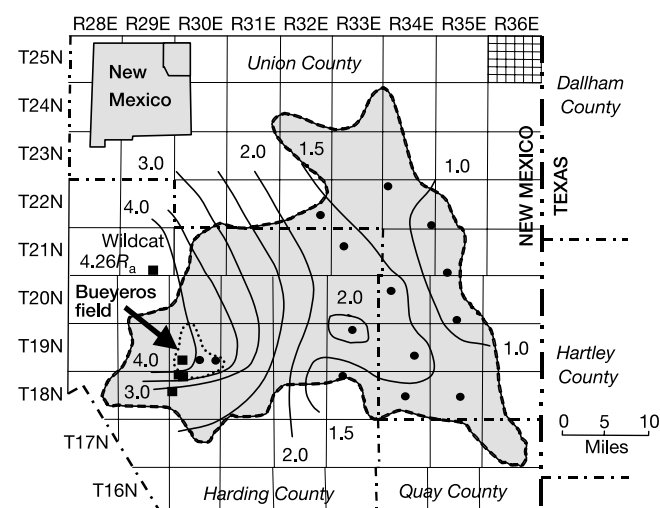


Figure 1 Study area. Filled circles, locations of samples from this study. Contours, $^3\text{He}/^4\text{He}$ ratios (expressed as R/R_a) across the field. Filled squares, locations of samples collected from non-producing wells and one wildcat well not reported here, but used in the $^3\text{He}/^4\text{He}$ contouring. The location of the historical Bueyeros field is also shown.

$^{20}\text{Ne}/^{22}\text{Ne}$ (where I is ^3He , ^4He , ^{40}Ar , ^{36}Ar or ^{84}Kr). The planes for $I = ^4\text{He}$ and ^{40}Ar are shown in Fig. 2. The equations for the best-fit planes are presented in Supplementary Information. This result shows that the same three endmembers (from crust, mantle and groundwater) can account for the noble gases in all samples.

Resolving the mantle Ne isotopic composition

Importantly, groundwater-derived (dissolved air) and crustal-derived noble gases are pre-mixed with only a small amount of variance²⁶ (Fig. 2). We show in Fig. 3a how mixing between an (air+crust) and a mantle endmember would generate a mixing wedge that defines the mantle Ne isotopic composition at its intersection with the air–MORB mixing line^{2,27}. The Bravo dome Ne isotopic data are shown in Fig. 3b, and uniquely intersect the MORB–air line at $^{20}\text{Ne}/^{22}\text{Ne} = 12.20 \pm 0.05$ and $^{21}\text{Ne}/^{22}\text{Ne} = 0.0558 \pm 0.0008$. Details of the statistical analysis of the robustness of this model and fit can be found in Supplementary Information. The identification of the mantle Ne isotopic composition by using Ne isotopes alone provides a more robust estimate than other techniques that rely *a priori* on assumed mantle isotopic or elemental abundance information. These mantle $^{20}\text{Ne}/^{22}\text{Ne}$ and

$^{21}\text{Ne}/^{22}\text{Ne}$ isotopic values can, in the simplest case, be input into the equations for the planes defined by the data, and then the mantle $I/^{22}\text{Ne}$ determined. This in turn enables the mantle $^3\text{He}/^4\text{He}$, $^{40}\text{Ar}/^{36}\text{Ar}$ and $I/^{36}\text{Ar}$ ratios also to be defined. In the Methods section, we give a more detailed discussion of the resolution of these mantle values, which are presented in Table 2.

Mantle $^3\text{He}/^4\text{He}$, $^{40}\text{Ar}/^{36}\text{Ar}$ and elemental ratios

One of the key questions regarding interpretation of the Ne isotopic result is whether or not the mantle fluids from this continental setting are representative of the convecting mantle. This issue is addressed in detail by Reid and Graham²⁸, who conclude from a Nd and He isotope study that the nearby Late Cenozoic Raton–Clayton basalts, some of which extrude over the field, are dominated by a depleted (convecting) mantle signature. Examination of the range of mantle $^3\text{He}/^4\text{He}$, $^{40}\text{Ar}/^{36}\text{Ar}$ and noble gas elemental abundance patterns (Table 2, Fig. 4) also suggest a well-gas volatile source little different to that found at mid-ocean ridges. Resolved mantle $^3\text{He}/^4\text{He}$ ratios in individual samples range from $5.4R_a$ to $7.4R_a$ (see Methods). This range is indistinguishable from $7.04 \pm 0.25R_a$ measured in local xenoliths²⁸, and is also within the average convecting mantle range of $8.75 \pm 2.14R_a$. The resolved well-gas mantle $^{40}\text{Ar}/^{36}\text{Ar}$ of between 37,000 to 55,200 is significantly higher than the value of 25,000 estimated from bulk analysis of the volatile-rich basalt glass IID43 ‘popping rock’ used as a reference sample for many noble gas models of the mantle system²⁷. However, variability of $^{40}\text{Ar}/^{36}\text{Ar}$ in bulk basalt glass samples can occur even for the same $^{20}\text{Ne}/^{22}\text{Ne}$ value¹⁶, and is probably related to atmospheric contamination of whole-rock analyses¹⁸. Laser decrepitation of individual fluid inclusions in the popping-rock sample record values of up to $^{40}\text{Ar}/^{36}\text{Ar} = 64,000 \pm 15,000$ (ref. 29). Studies of other MORB samples report values as high as $42,400 \pm 9,700$ (ref. 30). The question has been raised as to whether even the highest observed values are subject to contamination²⁹. Well-gas samples are not subject to the same mechanism of air contamination as basalt glasses, and are corrected for the air component (see Methods). The resolved range of mantle $^{40}\text{Ar}/^{36}\text{Ar}$ in the well gases therefore confirms the highest measured values, and defines (to our knowledge, for the first time) the $^{40}\text{Ar}/^{36}\text{Ar}$ limit for the convecting mantle (Table 2).

Confirmation of the high mantle $^{40}\text{Ar}/^{36}\text{Ar}$ has an important consequence. Bulk analysis popping-rock data are used to determine the convecting mantle noble gas abundance pattern, which is normalized to ^{36}Ar . On the basis of the results of this study, the mantle ^{36}Ar from the popping-rock whole-rock analyses has been overestimated by a factor of up to ~ 2 , and is clearly subject to some degree of air contamination. Whereas this correction can be made

Table 1 Noble gas results

Sample	$^3\text{He}/^4\text{He}$ (R/R_a)	Error	$^{20}\text{Ne}/^{22}\text{Ne}$	Error	$^{21}\text{Ne}/^{22}\text{Ne}$	Error	$^{40}\text{Ar}/^{36}\text{Ar}$	Error	^4He (10^{-5})	Error (10^{-6})	^{20}Ne (10^{-9})	Error (10^{-11})	^{40}Ar (10^{-5})	Error (10^{-7})	^{84}Kr (10^{-10})	Error (10^{-12})
BD01	1.670	0.008	10.664	0.034	0.05623	0.00031	10,723	393	9.44	1.1	1.69	1.7	3.03	2.9	1.01	3.6
BD02	0.764	0.004	9.955	0.035	0.05006	0.00032	4,660	59	41.5	5.0	7.00	7.1	6.52	5.9	5.04	13
BD03	0.896	0.004	10.014	0.012	0.05152	0.00014	5,346	92	33.1	4.0	5.21	4.7	5.36	4.2	3.24	8.0
BD04	1.611	0.008	10.585	0.043	0.05409	0.00035	9,911	237	9.61	1.1	1.81	1.9	2.86	2.4	1.03	2.5
BD05	0.965	0.005	9.930	0.014	0.05255	0.00024	5,412	56	27.0	3.3	4.46	4.0	5.38	4.2	3.33	7.4
BD06	1.503	0.008	10.488	0.038	0.05612	0.00043	9,213	206	12.0	1.4	2.02	2.0	3.50	2.8	1.35	3.1
BD07	2.104	0.011	11.201	0.049	0.05416	0.00041	10,955	387	7.81	0.95	1.80	2.1	2.80	2.5	0.900	3.5
BD08	1.143	0.006	10.214	0.027	0.05778	0.00037	6,650	90	16.1	1.9	2.64	2.5	3.96	3.3	2.00	4.7
BD09	1.724	0.009	10.735	0.051	0.05779	0.00054	ND		9.80	1.2	1.80	2.1	ND		ND	
BD10	1.104	0.006	10.196	0.021	0.05371	0.00029	6,726	106	19.9	2.4	3.08	2.9	3.96	3.0	2.02	4.3
BD11	3.784	0.019	11.880	0.047	0.05647	0.00044	21,560	1,591	3.91	0.47	1.03	1.3	2.41	3.5	0.455	1.8
BD12	3.627	0.018	ND		ND		21,019	1,279	4.15	0.50	ND		2.42	2.9	0.467	2.3
BD12 _{repeat}	3.634	0.018	11.603	0.061	0.05373	0.00024	22,612	3,061	4.13	0.50	1.20	1.5	2.40	3.8	0.490	2.3
BD13	1.318	0.007	10.250	0.046	0.05792	0.00045	7,735	277	15.3	1.8	2.40	2.8	3.82	3.8	1.92	4.9
BD14	1.413	0.007	10.543	0.125	0.05825	0.00023	8,517	645	11.5	1.4	1.79	3.2	3.07	2.7	1.23	3.3
WBD2	4.070	0.022	11.617	0.178	0.05583	0.00139	21,452	460	3.76	0.282	1.09	2.4	ND		0.404	11
Bueyeros (ref. 24)	3.136	0.321	11.6	0.9	0.062	0.008	16,200	340	1.67	6.30	9.86	9.8	3.3	9.8	0.435	0.41

All values corrected for full procedural blanks equivalent to $(5 \pm 2) \times 10^{-9} \text{ cm}^3 \text{ STP } ^4\text{He}$, $(5 \pm 2) \times 10^{-11} \text{ cm}^3 \text{ STP } ^{20}\text{Ne}$, $(6 \pm 2) \times 10^{-8} \text{ cm}^3 \text{ STP } ^{40}\text{Ar}$, $(1.4 \pm 0.52) \times 10^{-11} \text{ cm}^3 \text{ STP } ^{84}\text{Kr}$. Blank Ne and Ar isotopic ratios are atmospheric, while the blank $^3\text{He}/^4\text{He} = (1.397 \pm 0.007)R_a$. Concentrations in $\text{cm}^3 \text{ STP cm}^{-3}$. Errors are 1σ . ND, not determined.

with confidence to $\text{He}/^{36}\text{Ar}$ and $\text{Ne}/^{36}\text{Ar}$ ratios, the mechanism of air addition to basalt glasses remains poorly understood¹⁸, and the possibility that air Kr and Xe contamination has occurred in the basalt glass reference sample cannot be discounted. The mantle noble gas abundance pattern derived from well gases, with the air component unambiguously removed, therefore provides a new and critical data resource (Table 2), and is little different to the new $^{36}\text{Ar}_{\text{air}}$ -corrected popping rock (Fig. 4) (also see Supplementary Information).

The origin of resolved $^{20}\text{Ne}/^{22}\text{Ne}$

There are four possible reasons why the resolved $^{20}\text{Ne}/^{22}\text{Ne}$ of the well gases is lower than the solar value of 13.8: (1) enhanced ^{22}Ne nucleogenic production in the mantle is reducing a ratio that was originally solar; (2) solar–air mixing; (3) enrichment in the heavier isotopes of a reservoir residual from a mass-fractionating process such as diffusion; or (4) a trapped component (Ne-B) in accreting material^{16,17} is dominating the Ne inventory of the mantle source. To generate a mantle $^{20}\text{Ne}/^{22}\text{Ne} = 12.2$ – 12.5 and $^{21}\text{Ne}/^{22}\text{Ne} = 0.0558$ from solar values by nucleogenic production requires a mantle $^{21}\text{Ne}/^{22}\text{Ne}$ production ratio of 0.28, some 166 times lower than the predicted value^{2,31}. Crustal gas values are lower than those predicted for an homogeneous system by a factor of 5–7 because of preferential U + Th siting with the dominant ^{22}Ne -producing target element (fluorine)³¹. However, in the mantle only about one-third of the nucleogenic ^{22}Ne is produced by α -particle interaction with fluorine, requiring this route to be enhanced by a factor of ~ 500 (ref. 2). Element siting effects are unlikely to be able to account for the difference (more than two orders of magnitude) required.

Reduction of solar $^{20}\text{Ne}/^{22}\text{Ne} = 13.8$ to 12.5 by air-recycling into the mantle requires a 30% air Ne ($^{20}\text{Ne}/^{22}\text{Ne} = 9.8$) addition. We do not envisage any process that could introduce an elementally unfractionated air component into the mantle. Most systems, including sediments³² and water, are variably enriched in heavy noble gases relative to modern air. If we take deep sea water as an example, a 30% air Ne addition to the mantle would result in 310%,

220% and 130% of the Ar, Kr and Xe respectively in the popping rock being air-derived. It is shown above that because of air contamination Kr and Xe are overestimated in the popping-rock sample, and this estimate then represents a minimum. It follows that the addition of air-Ne with appropriate heavy noble gas enrichment would result in mantle $^{38}\text{Ar}/^{36}\text{Ar}$, Kr and Xe isotopic compositions dominated by air. $^{129}\text{Xe}/^{130}\text{Xe}$ ratios in excess of the air value found in both MORB popping rock³³ and Bravo dome well gas^{23–25} eliminates this possibility.

Preferential loss of lighter isotopes to reduce the $^{20}\text{Ne}/^{22}\text{Ne}$ from 13.8 to 12.5 by mass-dependent diffusion can be modelled as a simple reduced-mass Rayleigh fractionation. This would require an 80% loss of ^{22}Ne from an originally solar mantle reservoir. The same loss would reduce the pre-deuterium burning $^3\text{He}/^{22}\text{Ne}$ ratio from a solar value of 1.5 to 0.5. This is in the wrong direction to produce the resolved $^3\text{He}/^{22}\text{Ne} = 2.6$ – 2.8 in the well gas, which are minimum values in the case of phase fractionation effects during magmatic degassing (see Supplementary Information). At $^{20}\text{Ne}/^{22}\text{Ne} = 12.5$, $^3\text{He}/^{22}\text{Ne}$ in popping rock is even higher at 4.9 (Table 2), requiring an initial mantle $^3\text{He}/^{22}\text{Ne}$ of 15, an order of magnitude greater than the solar nebula value. Similar arguments apply when considering the similarity of the resolved mantle $^{84}\text{Kr}/^{22}\text{Ne}$ ratio in both popping rock and well gases (Table 2). We discount mass-dependent fractionation.

The value of the convecting mantle resolved in this study is indistinguishable from the Ne-B component found in solar-wind-irradiated meteoritic material. It would be fortuitous if the convecting mantle had evolved a Ne isotopic composition by either ancient (pre- ^{129}I , ^{244}Pu decay) air-addition or nucleogenic production to this same value, and we have ruled out the most viable processes that could do this from their predicted effect on the other noble gases. The most probable explanation for low $^{20}\text{Ne}/^{22}\text{Ne}$ in the well-gas mantle source is that it originated as a trapped component (Ne-B) implanted by solar corpuscular irradiation of accreting material^{16,17}. To preserve this value, subsequent atmosphere recycling, nucleogenic or other non-Ne-B neon admixtures to the mantle must have been minimal.

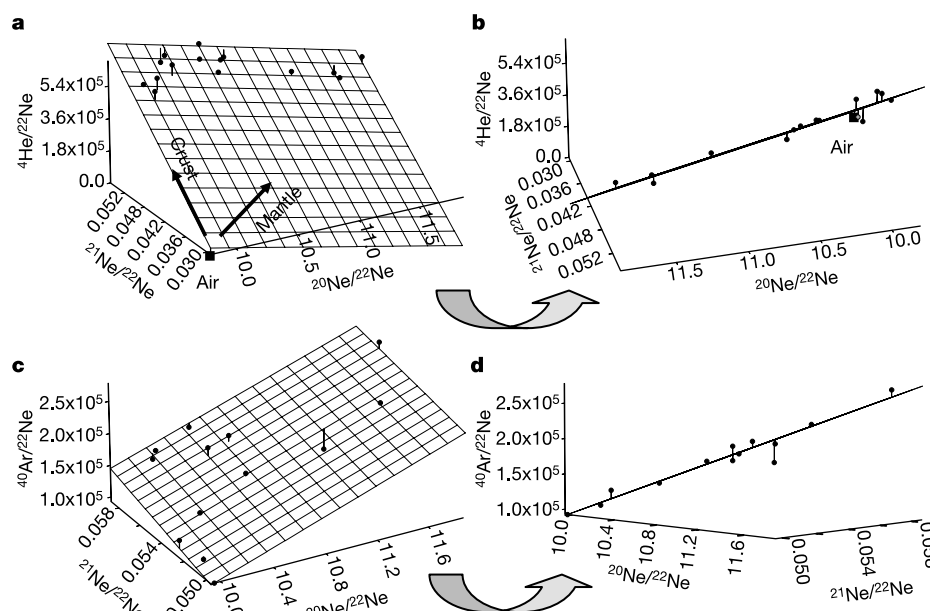


Figure 2 Correlation of measured $^{21}\text{Ne}/^{22}\text{Ne}$, $^{20}\text{Ne}/^{22}\text{Ne}$ with $^4\text{He}/^{22}\text{Ne}$ and $^{40}\text{Ar}/^{22}\text{Ne}$. **a, c**, Planes of best fit; **b, d**, rotations of the graph to view the plane edge-on. **a**, The small variance in the crust+air component suggests that the air- and crustal-derived noble gases are premixed in the groundwater system²⁶ before variable interaction with the

magmatic fluid. **c**, The $^{40}\text{Ar}/^{22}\text{Ne}$ fit to a plane again passes close to the air endmember (not shown). **b, d**, The good fit to a plane shows that the elemental abundance and isotopic data can be described by the same three endmember components for all samples.

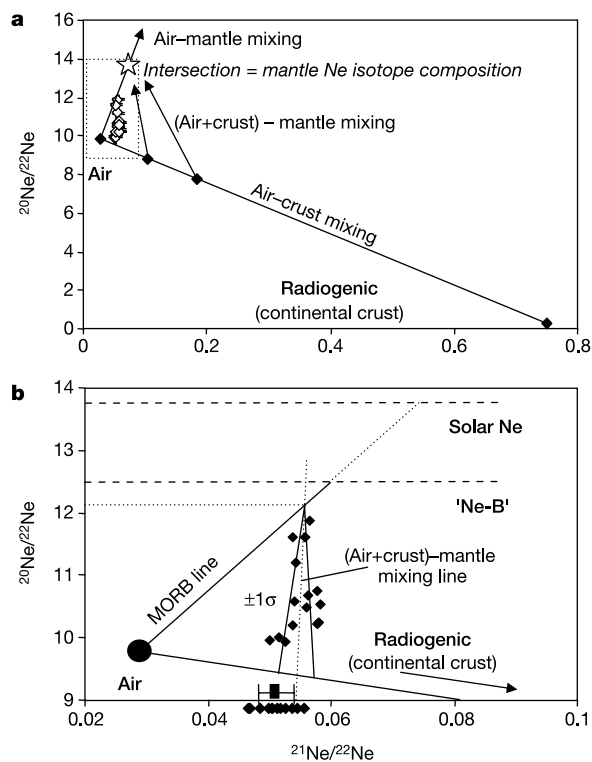


Figure 3 Intersection of a two-component air+crust mixture with the MORB air–mantle mixing line defines the mantle Ne isotopic endmember. **a**, Diagram showing different possible mixing lines. **b**, Bravo dome data projected onto the $^{20}\text{Ne}/^{22}\text{Ne}$ versus $^{21}\text{Ne}/^{22}\text{Ne}$ plane, showing a pseudo-two-component mix between air+crust and mantle components. The scatter from a single mixing line is because of a small amount of variance in the air+crust mixing ratio. The average and 1σ variance of the resolved air+crust component (see Supplementary Information) is shown on the x axis. The data intersect the MORB air–mantle line^{2,27} at $^{20}\text{Ne}/^{22}\text{Ne} = 12.20 \pm 0.05$.

Convection models and the origin of volatiles in the mantle

MORB has higher and distinct $^3\text{He}/^{22}\text{Ne}$ compared with Hawaii³⁴, Iceland^{35–37} or Kola¹⁹. Fluxing models therefore need to have different residence times for He and Ne in the convecting mantle. This could be caused by differential extraction at mid-ocean ridges. Alternatively, preferential transferral of He relative to Ne from the proposed volatile-rich source needs to be invoked by this form of model. Although these processes cannot be simply ruled out, they provide an additional level of complexity when justifying the flux models. The observation that the Iceland plume source has a $^{20}\text{Ne}/^{22}\text{Ne}$ ratio indistinguishable from solar (13.75 ± 0.32 ; ref. 37) contrasts markedly with the Ne isotopic ratio of the mantle source resolved in this work. Although the Iceland measurement is the subject of debate^{16,18}, more recent data from the Kola peninsula reports reproducible plume source $^{20}\text{Ne}/^{22}\text{Ne} > 13.0 \pm 0.2$ (ref. 19). This is again consistent with a solar rather than a Ne-B

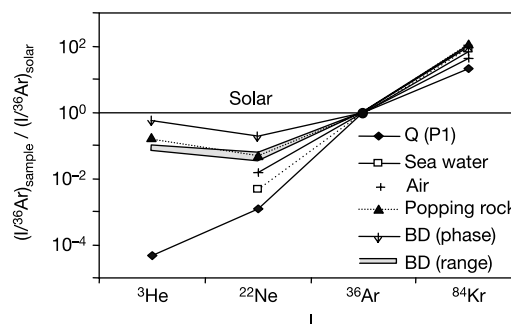


Figure 4 Mantle noble gas isotopes relative to ^{36}Ar normalized to the solar abundance, relative to ^{36}Ar (Table 2). The range of values derived from the Bush dome study is shown as a shaded area, BD (range), and compared with the popping-rock sample, corrected for the original literature ^{36}Ar overestimate (see text). With the exception of a small ^3He deficit in the well-gas system, the elemental abundance pattern of popping rock and the well gases are indistinguishable. Also shown is the limit estimated for phase fractionation of the magmatic gases, BD (phase) (see Supplementary Information), air, sea water and the Q (P1) trapped component in primitive meteorites⁵⁰.

origin for Ne in the plume source.

An average noble gas concentration in the convecting mantle that is higher by a factor of ~ 3.5 would remove the need for either a ^3He flux into the convecting mantle or a hidden ^{40}Ar -rich reservoir, and the heat–helium imbalance could then be accounted for by temporal variance in the ocean-ridge helium flux within a whole-mantle convective regime³⁸. This has been called the ‘zero paradox’ model concentration. This would allow the decoupling of the plume and convecting-mantle volatile systems suggested by the $^3\text{He}/^{22}\text{Ne}$ and Ne isotopic difference highlighted here. But recent constraints^{39–41} on average convecting mantle Nb/ CO_2 , $\text{CO}_2/^3\text{He}$ and Nb make it unlikely that the ‘zero paradox’ ^3He concentration is reached. Uncertainty within these values means that the ^3He concentration may nevertheless be higher than that used by current flux models, reducing the noble gas flux required from a deep mantle source. As there is no obvious process capable of reducing a solar Ne isotopic ratio to the Ne isotopic value observed in the well gases, the source of the Kola and perhaps the Icelandic plume material cannot provide the noble gases in the convecting mantle. We therefore argue that any solar Ne heterogeneity is only a small component of a Ne-B dominated deep mantle source.

The neon isotope heterogeneity of the mantle also has profound implications for the processes and timing of Earth accretion. A significant fraction of terrestrial accretion took place when the solar nebula was still present in the first 10 Myr of the accretionary process. During this period, planetary bodies the size of Mars formed and differentiated, indicated independently by dynamical simulations⁴² and the extinct radioactivity systematics of SNC meteorites and other differentiated bodies^{43–45}. Thus, solar-nebula gas could have been trapped directly during accretion and compaction, or indirectly during magmatic episodes and equilibration with

Table 2 Resolved mantle elemental and isotopic ratios

	$^3\text{He}/^4\text{He}$	$^{20}\text{Ne}/^{22}\text{Ne}$	$^{40}\text{Ar}/^{36}\text{Ar}$	$^4\text{He}/^{21}\text{Ne}^*$	$^3\text{He}/^{22}\text{Ne}$	$^4\text{He}/^{40}\text{Ar}$	$^3\text{He}/^{36}\text{Ar}$	$^{22}\text{Ne}/^{36}\text{Ar}$	$^{84}\text{Kr}/^{36}\text{Ar}$
Bravo dome	$7.05R_a$	12.5	55,200	1.23×10^7	2.56	1.09	0.50	0.1953	0.0545
Bravo dome	$5.35R_a$	12.2	37,000	1.45×10^7	2.77	0.844	0.33	0.1202	0.0482
Popping rock†	$8.3R_a$	12.5	25,000–44,000*	1.68×10^7	4.90	$1.06^* - 1.52$	$0.40 (0.8)\S$	$0.0816 (0.163)\S$	$0.0284 (0.0568)\S$
Air	—	9.8	295.5	—	—	—	—	0.0530	0.0207
Sea water	—	9.8	295.5	—	—	—	—	0.0156	0.0405
Solar	—	13.8	—	—	$1.5\ddagger$	—	$5.1\ddagger$	3.4	5.05×10^{-4}

$^{21}\text{Ne}^*$ indicates mantle ^{21}Ne corrected for solar contribution².

*Ref. 29.

†Ref. 27.

‡Pre-deuterium burning, see ref. 19.

§Corrected for ^{36}Ar overestimate (see text).

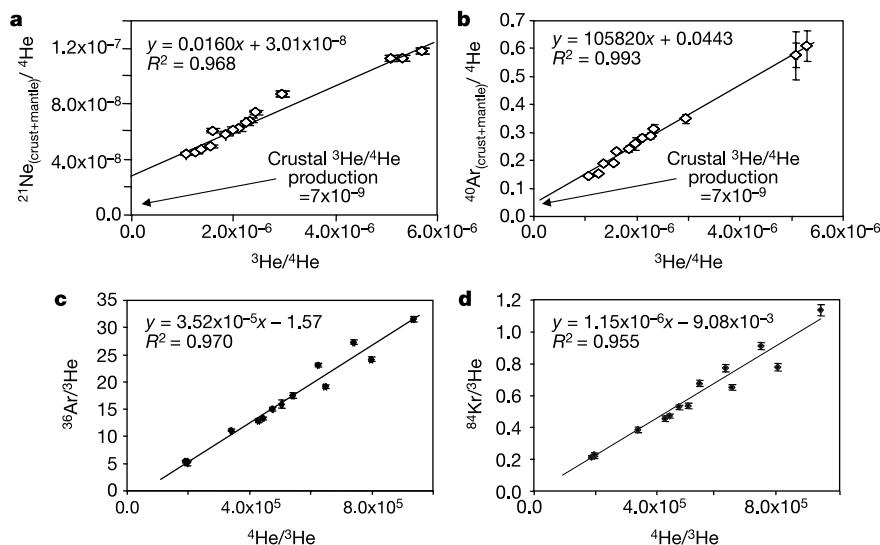


Figure 5 Resolving endmember compositions from simple mixing. Elemental ratios plotted against $^3\text{He}/^4\text{He}$ (**a**, **b**) or $^4\text{He}/^3\text{He}$ (**c**, **d**) are used to derive crustal and mantle endmember values (see Methods and Table 2). Air-corrected ^{21}Ne (using mantle $^{20}\text{Ne}/^{22}\text{Ne} = 12.5$) (**a**) and air-corrected ^{40}Ar (**b**) are normalized to ^4He to generate true

mantle–crust two-component mixing lines. Measured $^{36}\text{Ar}/^3\text{He}$ (**c**) and $^{84}\text{Kr}/^3\text{He}$ (**d**) versus $^4\text{He}/^3\text{He}$ define pseudo-two-component mixing lines between crust+air and mantle endmembers. All y -axis error bars are 1σ confidence. Errors in $^3\text{He}/^4\text{He}$ (Table 1) or $^4\text{He}/^3\text{He}$ are within the plotted symbol.

massive early planetary atmospheres⁴⁶. It has also been shown that under such circumstances a significant amount of noble gases could be partitioned into core-forming phases⁴⁷. Gas loss processes could also occur, for example, during collision with large impactors⁴⁶. In the case of the solid Earth, extensive volatile loss up to 80 Myr after accretion is shown by the relative proportions of the Xe daughter products of extinct ^{129}I and ^{244}Pu (ref. 2). Late infall of solar-corpiscular-irradiated material could then form a significant volatile source for the degassed planetary bodies, with entrainment into the mantle possible in an early, dry subducting system⁴⁸. The occurrence of solar-nebula gases in plumes is then evidence of either a remnant of the silicate mantle that preserves an accretionary volatile history distinct from that of the convecting mantle, or a portion of the mantle that has received its different volatile composition as a result of interaction with the core. The D'' density contrast with the convecting mantle may have helped to preserve the distinct character of this portion of the mantle over time⁴⁸. The results presented here, however, suggest that the bulk of the terrestrial silicate mantle He and Ne is dominated by solar-corpiscular-irradiated accretionary material. □

Methods

Resolving the mantle noble gas components

Resolution of the isotopic and elemental characteristics of the mantle noble gases depends first on the resolved mantle $^{20}\text{Ne}/^{22}\text{Ne}$ and $^{21}\text{Ne}/^{22}\text{Ne}$ values. A magmatic fluid interacting with a fluid at a near-constant air+crust mixing ratio defines a unique (air+crust)–mantle mixing wedge (Fig. 3). The intersection of the data wedge with the air–MORB line at $^{20}\text{Ne}/^{22}\text{Ne} = 12.20 \pm 0.05$ is, however, lower than the accepted Ne–B value of 12.5. As we discuss, there are few viable mechanisms to reduce the $^{20}\text{Ne}/^{22}\text{Ne}$ ratio from solar to lower values, and we must consider the possibility that the choice of the air–MORB line may introduce some uncertainty. We therefore consider $^{20}\text{Ne}/^{22}\text{Ne}$ mantle values along the mixing line up to 12.5 (Fig. 3), providing a sensitivity test for the resolved mantle components. Although input of the mantle Ne isotopic values into the equations for the respective equations for a plane provides the simplest conceptual method of determining the relationship between Ne and the other mantle-derived noble gases, care must be taken with this approach as the plane fits are neither error-weighted nor have an error assessment of the extrapolated values. We show here how the Ne isotopes are used to allow a full error assessment of resolved components.

Crustal noble gas isotopic production ratios, like those in air, are well defined³¹. The mantle-, air- and crustal-Ne isotope endmembers then define the proportion of these components in each individual sample⁴⁹ (Supplementary Tables 1 and 2), where, for example:

$$^{21}\text{Ne}_{\text{total}} = ^{21}\text{Ne}_{\text{air}} + ^{21}\text{Ne}_{\text{crust}} + ^{21}\text{Ne}_{\text{mantle}} \quad (1)$$

Resolved $^{21}\text{Ne}_{\text{air}}$ is subtracted from $^{21}\text{Ne}_{\text{total}}$ to leave ‘air-corrected’ ^{21}Ne , $^{21}\text{Ne}_{(\text{crust+mantle})}$. A plot of $^{21}\text{Ne}_{(\text{crust+mantle})}/^4\text{He}$ against $^3\text{He}/^4\text{He}$, with a negligible air-helium component, then develops a simple two-component mixing line (Fig. 5a). Extrapolation to the crustal $^3\text{He}/^4\text{He}$ endmember value ($^3\text{He}/^4\text{He}_{\text{crust}} = 0.007R_a$)³¹ defines the local $^4\text{He}/^{21}\text{Ne}$ crustal input value, $^4\text{He}/^{21}\text{Ne}_{\text{crust}}$, to be between $(3.47 \pm 0.24) \times 10^7$ and $(3.30 \pm 0.23) \times 10^7$ for $^{20}\text{Ne}/^{22}\text{Ne} = 12.2$ and 12.5 (Fig. 5a), respectively. These values are indistinguishable, and illustrate the insensitivity of the resolved crustal components to the choice of the mantle Ne isotopic endmember value.

The crustal ^3He and ^4He contribution to each sample can then be calculated from $^{21}\text{Ne}_{\text{crust}}$, $^3\text{He}/^4\text{He}_{\text{crust}}$ and $^4\text{He}/^{21}\text{Ne}_{\text{crust}}$. This crustal ^3He and ^4He component can be subtracted from the observed crust/mantle helium mixture to determine the mantle $^3\text{He}/^4\text{He}$ endmember for each sample. Three samples give a mantle component $^3\text{He}/^4\text{He}$ determination with propagated errors of $<50\%$. For $^{20}\text{Ne}/^{22}\text{Ne} = 12.2$ and 12.5, samples BD07, BD11 and BD12 give crust-corrected $^3\text{He}/^4\text{He}$ values of $5.18 \pm 0.76R_a$ to $7.42 \pm 1.61R_a$, $5.29 \pm 0.42R_a$ to $7.11 \pm 0.77R_a$ and $5.58 \pm 0.52R_a$ to $7.69 \pm 1.00R_a$, with an error-weighted average of $5.35 \pm 0.36R_a$ to $7.4 \pm 0.5R_a$, for the magmatic $^3\text{He}/^4\text{He}$ endmember. These compare with $4.9\text{--}6.0R_a$ derived from the full data set defined planes (Supplementary Information). Extrapolation of the $^{21}\text{Ne}_{(\text{crust+mantle})}/^4\text{He}$ mixing line to mantle $^3\text{He}/^4\text{He}$ values defines the range of $^{21}\text{Ne}/^4\text{He}_{\text{mantle}}$ values resolved by this data set.

Crustal and mantle $^{40}\text{Ar}/^{36}\text{Ar}$ ratios are both large, 6.1×10^7 (ref. 31) and $>2.5 \times 10^5$ respectively, compared with the air ratio. This allows ^{40}Ar in excess of $^{40}\text{Ar}_{\text{air}}$ $^{40}\text{Ar}_{(\text{crust+mantle})}$ to be calculated, where:

$$^{40}\text{Ar}_{(\text{crust+mantle})} = ^{40}\text{Ar}_{\text{total}} [1 - (^{40}\text{Ar}/^{36}\text{Ar})_{\text{air}} / (^{40}\text{Ar}/^{36}\text{Ar})_{\text{sample}}] \quad (2)$$

$^{40}\text{Ar}_{(\text{crust+mantle})}/^4\text{He}$ plotted against $^3\text{He}/^4\text{He}$ then also generates a simple two-component mixing line (Fig. 5b). Extrapolation to the mantle $^3\text{He}/^4\text{He}$ values gives the range of resolved $^{40}\text{Ar}/^4\text{He}_{\text{mantle}}$ ratios.

$^{36}\text{Ar}_{\text{total}}$ and $^{84}\text{Kr}_{\text{total}}$ are treated as two-component mixtures of crust+air and mantle. Although there is some variance in the crust+air mixing ratio (Fig. 2), plots of $^{36}\text{Ar}_{\text{total}}/^3\text{He}$ and $^{84}\text{Kr}_{\text{total}}/^3\text{He}$ versus $^4\text{He}/^3\text{He}$ produce pseudo-two-component mixing lines that converge towards the mantle $^4\text{He}/^3\text{He}$ (Fig. 5c, d). The convergence of data and high correlation coefficients provide confidence in this approach. Extrapolation to the mantle $^4\text{He}/^3\text{He}$ values then gives the range of $^{36}\text{Ar}/^3\text{He}_{\text{mantle}}$ and $^{84}\text{Kr}/^3\text{He}_{\text{mantle}}$ respectively. Combining the mantle $^3\text{He}/^4\text{He}$ with $^{21}\text{Ne}/^4\text{He}_{\text{mantle}}$, $^{40}\text{Ar}/^4\text{He}_{\text{mantle}}$, $^{36}\text{Ar}/^3\text{He}_{\text{mantle}}$ and $^{84}\text{Kr}/^3\text{He}_{\text{mantle}}$ provides the range of isotopic and elemental abundance ratios presented in Table 2.

Received 9 July; accepted 4 November 2004; doi:10.1038/nature03182.

- Allège, C. J., Staudacher, T., Sarda, P. & Kurz, M. Constraints on evolution of Earth's mantle from rare gas systematics. *Nature* **303**, 762–766 (1983).
- Porcelli, D. & Ballentine, C. J. Models for the distribution of terrestrial noble gases and evolution of the atmosphere. *Rev. Min. Geochem.* **47**, 411–480 (2002).
- Porcelli, D. & Wasserburg, G. J. Mass transfer of helium, neon, argon, and xenon through a steady-state upper mantle. *Geochim. Cosmochim. Acta* **59**, 4921–4937 (1995).
- Hart, R., Dymond, J. & Hogan, L. Preferential formation of the atmosphere-sialic crust system from the upper mantle. *Nature* **278**, 156–159 (1979).
- O’Nions, R. K. & Oxburgh, E. R. Heat and helium in the Earth. *Nature* **306**, 429–431 (1983).
- Davies, G. F. & Richards, M. A. Mantle convection. *J. Geol.* **100**, 151–206 (1992).
- Van der Hilst, R. D., Widiyantoro, S. & Engdahl, E. R. Evidence for deep mantle circulation from

- global tomography. *Nature* **386**, 578–584 (1997).
8. Van Keken, P. E. & Ballentine, C. J. Dynamical models of mantle volatile evolution and the role of phase transitions and temperature-dependent rheology. *J. Geophys. Res.* **B 104**, 7137–7151 (1999).
9. Kellogg, L. H., Hager, B. H. & Van der Hilst, R. D. Compositional stratification in the deep mantle. *Science* **283**, 1881–1884 (1999).
10. Becker, T. W., Kellogg, J. B. & O'Connell, R. J. Thermal constraints on the survival of primitive blobs in the lower mantle. *Earth Planet. Sci. Lett.* **171**, 351–365 (1999).
11. Vidale, J. E., Schubert, G. & Earle, P. S. Unsuccessful initial search for a midmantle chemical boundary with seismic arrays. *Geophys. Res. Lett.* **28**, 859–862 (2001).
12. Bercovici, D. & Karato, S. Whole-mantle convection and the transition-zone water filter. *Nature* **425**, 39–44 (2003).
13. Wieler, R. Noble gases in the solar system. *Rev. Min. Geochem.* **47**, 21–70 (2002).
14. Kallenbach, R. *et al.* Isotopic composition of solar wind neon measured by CELIAS/MTOF on board SOHO. *J. Geophys. Res.* **A 102**, 26895–26904 (1997).
15. Black, D. C. On the origin of trapped helium, neon, and argon isotopic variations in meteorites — I Gas-rich meteorites, lunar soil and breccia. *Geochim. Cosmochim. Acta* **36**, 347–375 (1972).
16. Trieloff, M., Kunz, J., Clague, D. A., Harrison, D. & Allègre, C. J. The nature of pristine noble gases in mantle plumes. *Science* **288**, 1036–1038 (2000).
17. Trieloff, M., Kunz, J. & Allègre, C. J. Noble gas systematics of the Réunion mantle plume source and the origin of primordial noble gases in Earth's mantle. *Earth Planet. Sci. Lett.* **200**, 297–313 (2002).
18. Ballentine, C. J., Porcelli, D. & Wieler, R. Technical comment on 'Noble gases in mantle plumes' by Trieloff *et al.* (2000) and reply. *Science* **291**, 2269, doi:10.1126/science.291.5512.2269a (2001).
19. Yokochi, R. & Marty, B. A determination of the neon isotopic composition of the deep mantle. *Earth Planet. Sci. Lett.* **225**, 77–88 (2004).
20. Sherwood Lollar, B., Ballentine, C. J. & O'Nions, R. K. The fate of mantle-derived carbon in a continental sedimentary basin: Integration of C/He relationships and stable isotope signatures. *Geochim. Cosmochim. Acta* **61**, 2295–2307 (1997).
21. Ballentine, C. J., Schoell, M., Coleman, D. & Cain, B. A. 300-Myr-old magmatic CO₂ in natural gas reservoirs of the west Texas Permian basin. *Nature* **409**, 327–331 (2001).
22. Broadhead, R. F. Carbon dioxide in northeast New Mexico. *West Tex. Geol. Soc. Bull.* **32**, 5–8 (1993).
23. Staudacher, T. Upper mantle origin for Harding County well gases. *Nature* **325**, 605–607 (1987).
24. Phinney, D., Tennyson, J. & Frick, U. Xenon in CO₂ well gas revisited. *J. Geophys. Res.* **B 83**, 2313–2319 (1978).
25. Caffee, M. W. *et al.* Primordial noble gases from Earth's mantle: Identification of a primitive volatile component. *Science* **285**, 2115–2118 (1999).
26. Ballentine, C. J. & Sherwood Lollar, B. Regional groundwater focusing of nitrogen and noble gases into the Hugoton-Panhandle giant gas field, USA. *Geochim. Cosmochim. Acta* **66**, 2483–2497 (2002).
27. Moreira, M., Kunz, J. & Allègre, C. J. Rare gas systematics in popping rock: isotopic and elemental compositions in the upper mantle. *Science* **279**, 1178–1181 (1998).
28. Reid, M. R. & Graham, D. W. Resolving lithospheric and sub-lithospheric contributions to helium isotope variations in basalts from the southwestern US. *Earth Planet. Sci. Lett.* **144**, 213–222 (1996).
29. Burnard, P., Graham, D. & Turner, G. Vesicle specific noble gas analyses of popping rock: Implications for primordial noble gases in Earth. *Science* **276**, 568–571 (1997).
30. Marty, B. & Humbert, F. Nitrogen and argon isotopes in oceanic basalts. *Earth Planet. Sci. Lett.* **152**, 101–112 (1997).
31. Ballentine, C. J. & Burnard, P. G. Production, release and transport of noble gases in the continental crust. *Rev. Min. Geochem.* **47**, 481–538 (2002).
32. Torgersen, T. & Kennedy, B. M. Air-Xe enrichments in Elk Hills oil field gases: role of water in migration and storage. *Earth Planet. Sci. Lett.* **167**, 239–253 (1999).
33. Kunz, J., Staudacher, T. & Allègre, C. J. Plutonium-fission xenon found in Earth's mantle. *Science* **280**, 877–880 (1998).
34. Honda, M. & McDougall, I. Primordial helium and neon in the Earth—a speculation on early degassing. *Geophys. Res. Lett.* **25**, 1951–1954 (1998).
35. Dixon, E. T., Honda, M., McDougall, I., Campbell, I. H. & Sigurdsson, I. Preservation of near-solar neon isotopic ratios in Icelandic basalts. *Earth Planet. Sci. Lett.* **180**, 309–324 (2000).
36. Moreira, M., Breddam, K., Curtice, J. & Kurz, M. D. Solar neon in the Icelandic mantle: new evidence for an undegassed lower mantle. *Earth Planet. Sci. Lett.* **185**, 15–23 (2001).
37. Harrison, D., Burnard, P. & Turner, G. Noble gas behaviour and composition in the mantle: constraints from the Iceland Plume. *Earth Planet. Sci. Lett.* **171**, 199–207 (1999).
38. Ballentine, C. J., van Keken, P. E., Porcelli, D. & Hauri, E. H. Numerical models, geochemistry and the zero paradox noble-gas mantle. *Phil. Trans. R. Soc. Lond. A* **360**, 2611–2631 (2002).
39. Saal, E. S., Hauri, E. H., Langmuir, C. H. & Perfit, M. R. Vapour undersaturation in primitive mid-ocean-ridge basalt and the volatile content of Earth's upper mantle. *Nature* **419**, 451–455 (2002).
40. Marty, B. & Tolstikhin, I. N. CO₂ fluxes from mid-ocean ridges, arcs and plumes. *Chem. Geol.* **145**, 233–248 (1998).
41. Su, Y. & Langmuir, C. H. *Global MORB Chemistry Compilation at the Segment Scale* (Department of Earth and Environmental Sciences, Columbia University, 2003; available at (<http://petdb.ldeo.columbia.edu/documentation/morbcompilation/>).
42. Wetherill, G. W. in *Origin of the Moon* (eds Hartmann, W. K., Phillips, R. J. & Taylor, G. I.) 519–555 (Oxford Univ. Press, Oxford, 1986).
43. Halliday, A. N., Wänke, H., Birk, J. L. & Clayton, R. N. The accretion, composition and early differentiation of Mars. *Space Sci. Rev.* **96**, 1–34 (2001).
44. Yin, Q. Z. *et al.* A short timescale for terrestrial planet formation from Hf-W chronometry of meteorites. *Nature* **418**, 949–952 (2002).
45. Kleine, T., Munker, C., Mezger, K. & Palme, H. Rapid accretion and early core formation on asteroids and the terrestrial planets from Hf-W chronometry. *Nature* **418**, 952–955 (2002).
46. Porcelli, D., Woolum, D. S. & Cassen, P. Deep Earth rare gases: initial inventories, capture from the solar nebula, and losses during Moon formation. *Earth Planet. Sci. Lett.* **193**, 237–251 (2001).
47. Porcelli, D. & Halliday, A. N. The core as a possible source of mantle helium. *Earth Planet. Sci. Lett.* **192**, 45–56 (2001).
48. Tolstikhin, I. & Hofmann, A. W. Early crust on top of the Earth's core. *Phys. Earth Planet. Inter.* (in the press).
49. Ballentine, C. J. & O'Nions, R. K. The nature of mantle neon contributions to Vienna Basin hydrocarbon reservoirs. *Earth Planet. Sci. Lett.* **113**, 553–567 (1992).
50. Ott, U. Noble gases in meteorites—trapped components. *Rev. Min. Geochem.* **47**, 71–100 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements Access and permission to sample was by permission of BP (the field is now owned by Oxy) and the Bravo dome field manager, D. Holcomb. Sampling from the West Bush dome was by permission of Amerada Hess. H. Baur provided laboratory support. We thank F. Albarède, D. Porcelli, A. Halliday, A. Hofmann, C. Hall, J. Gilmour, G. Holland, D. Murphy, R. Yokochi and I. Tolstikhin for discussions and critical comments that have improved this Article. This work was funded by the Zurich ETH and NERC.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.J.B. (chris.ballentine@manchester.ac.uk).

The PIN auxin efflux facilitator network controls growth and patterning in *Arabidopsis* roots

Ikram Bilou¹, Jian Xu^{1*}, Marjolein Wildwater^{1*}, Viola Willemsen^{1*}, Ivan Paponov², Jiří Friml³, Renze Heidstra¹, Mitsuhiro Aida¹, Klaus Palme² & Ben Scheres¹

¹Department of Molecular Genetics, Utrecht University, Padualaan 8, 3584CH Utrecht, The Netherlands

²Institut für Biologie II, Universität Freiburg, Schänzlestrasse 1, 79104, Freiburg, Germany

³Zentrum für Molekularbiologie der Pflanzen, Universität Tübingen, Auf der Morgenstelle 3, 72076 Tübingen, Germany

* These authors contributed equally to this work

Local accumulation of the plant growth regulator auxin mediates pattern formation in *Arabidopsis* roots and influences outgrowth and development of lateral root- and shoot-derived primordia. However, it has remained unclear how auxin can simultaneously regulate patterning and organ outgrowth and how its distribution is stabilized in a primordium-specific manner. Here we show that five *PIN* genes collectively control auxin distribution to regulate cell division and cell expansion in the primary root. Furthermore, the joint action of these genes has an important role in pattern formation by focusing the auxin maximum and restricting the expression domain of *PLETHORA* (*PLT*) genes, major determinants for root stem cell specification. In turn, *PLT* genes are required for *PIN* gene transcription to stabilize the auxin maximum at the distal root tip. Our data reveal an interaction network of auxin transport facilitators and root fate determinants that control patterning and growth of the root primordium.

In *Arabidopsis* root development, a distal auxin maximum correlates with pattern formation and the orientation and extent of cell division¹. Inhibition of polar auxin transport strongly affects these processes¹. The initiation of lateral roots and leaf primordia is also associated with changes in auxin transport^{2–4}. These observations point to polar auxin transport as a major factor in organ formation. Recent analyses of the *Arabidopsis* PIN proteins revealed their association with auxin maxima in distal domains of organ primordia^{2–4}. The *PIN* genes are thought to encode components of the auxin efflux machinery mediating polar auxin transport, as concluded from the polar localization of PIN membrane proteins and auxin uptake experiments^{5–11}. PIN proteins might participate directly in auxin transport or help in the assembly of other proteins with efflux activity such as the AtMDR/PGP proteins^{12,13}. Mutations in the ARF-GEF GNOM, required for vesicle transport of PIN1, also lead to developmental defects that resemble those caused by interfering with auxin transport^{14–16}.

All single *pin* mutants described so far display different weak phenotypes in primary roots^{4–8}, in contrast with *gnom* mutants and polar auxin transport inhibitor treatments. Here, we investigate the collective contribution of five *PIN* genes in the control of cell division and cell expansion during root outgrowth. Furthermore, we assess how the *PLETHORA* (*PLT*) genes, which encode auxin-inducible AP2-domain putative transcription factors necessary and sufficient for stem cell specification in the root primordium¹⁷, respond to and regulate *PIN* gene activity during pattern formation.

PIN protein localization is changed in *pin* mutants

The PIN proteins described so far are expressed in specific but overlapping regions of the root meristem^{6–8}. PIN1 mainly resides at the basal end of the vascular cells⁸ but weak PIN1 signals can be detected in the epidermis and the cortex (Fig. 1b). PIN2 localizes apically in epidermal and lateral root cap cells and predominantly basally in cortical cells⁶ (Fig. 1e, f). PIN3 is expressed without pronounced polarity in tiers two and three of the columella cells, at

the basal side of vascular cells and to the lateral side of pericycle cells of the elongation zone⁷ (Fig. 1h, i). PIN4 is detected around the quiescent centre and cells surrounding it, and localizes basally in provascular cells⁸ (Fig. 1k). PIN7 resides at lateral and basal membranes of provascular cells in the meristem and elongation zone, whereas in the columella cells it coincides with the PIN3 domain (Fig. 1j).

Auxin distribution appears to be altered in *pin* mutants^{1,8} and differences in auxin homeostasis affect PIN2 expression¹⁸. We therefore determined whether *pin* mutants have altered expression of remaining PIN proteins. In *pin3pin4pin7*, enhanced PIN1 protein was detected in lateral-basal membranes of the endodermis (Fig. 1c, d). Moreover, ectopic PIN2 protein was detected at the basal end of provascular cells that normally express PIN3 and PIN7 (Fig. 1g). PIN4 expression expands to tier three of the columella cells in *pin3* single mutants and *pin2pin3* double mutants with a membrane localization similar to PIN3 in the wild type (Fig. 1l and data not shown). In *pin3pin7*, PIN4 expands to the lateral root cap (Fig. 1m).

Our data show that defects in *pin* mutants can be masked by ectopic activity of the remaining *PIN* genes. A comprehensive mutant analysis is therefore necessary to uncover full gene function. To this end, we generated all mutant combinations for *PIN1*, *PIN2*, *PIN3*, *PIN4* and *PIN7*, which group together within the eight-member *PIN* gene family. We verified phenotypes in independent allelic combinations to exclude influences of background modifiers (Supplementary Table 1).

PIN genes control cell division zone size in the root meristem

Classical experiments revealed that externally added auxins can stimulate cell division¹⁹ but it is unknown whether internal auxin distribution regulates cell division in primordia. *PIN* genes are required for outgrowth of all organs^{2,3,20}, and we assessed whether this reflected their contribution to the *in vivo* regulation of cell division. In roots, oriented cell divisions accompanied by a low rate of cell expansion occur in the distal meristem zone (Fig. 2A).

pin1 and *pin2* single mutants display a slight reduction of root

length and root meristem size (Supplementary Fig. 2) whereas *pin3*, *pin4* and *pin7* single mutants only display subtle division defects in the quiescent centre and columella root cap⁸ (Supplementary Fig. 1d–h). Most double-mutant combinations show additive defects in orientation of cell division, root length and root meristem size (Supplementary Figs 1 and 2). However, *pin1pin2* and all triple and quadruple mutants containing *pin2* show more-than-additive reduction in root size and root meristem size (Fig. 2B, C and Supplementary Fig. 2a, b), suggesting that PIN2 plays a pivotal role in cell division control. The *pin1pin2* double mutant suggests that the role of PIN2 in meristem size control is masked in the *pin2* single mutant mainly by the activity of PIN1 in the PIN2 domain (Fig. 1b). PIN2 is a main component for mediating proximal (basipetal) auxin transport (Fig. 1e, f, n), which implies that basipetal transport to meristematic cells has a critical role in meristem length regulation. Consistent with the notion of auxin shortage in the meristem

zone, treatment with auxins restored meristem size of *pin1pin2* and *pin2pin3pin7* to that of wild type (data not shown). Our data substantiate physiological evidence indicating a role for basipetal auxin transport in root growth^{21,22} and identify control of cell division as a major factor in this process.

The extreme reduction in root meristem size in *pin2* mutant combinations is reminiscent of the phenotype obtained upon diphtheria toxin-mediated genetic ablation of root cap cells²³. In such plants we found high *DR5::GUS* accumulation in the provascular tissue (Fig. 2E, F), supporting the notion that lateral auxin redistribution does not occur in the absence of columella cells with laterally oriented PIN proteins.

Our data imply that basipetal transport and lateral redistribution of auxin are both critical for maintenance of the meristem zone. A potential mechanism to transport auxin to every cell in the meristem is indicated by PIN2 localization at basal membranes of cortex cells (Fig. 1e, f) and the expression domains of vascular PIN3 and PIN7, which suggest an ‘auxin reflux’ loop (Fig. 1h, j, n). We investigated auxin transport in the root by expressing the bacterial auxin biosynthetic enzyme IAAH under the *WOX5* (ref. 24) promoter, which allows induction of auxin biosynthesis in the quiescent centre (Fig. 2G) by external addition of the precursor IAM²⁵. After different time points of induction, we monitored auxin accumulation using the *DR5-GFP* reporter. In line with the observed polar orientation of the PIN proteins, enhanced auxin responses appear first in the columella region, subsequently in the lateral root cap and then in the epidermis; enhanced responses in provascular strands only emerge at later stages (Fig. 2H, a–d). Epidermal and provascular auxin responses could be reduced by the polar transport inhibitor N-naphthylphthalamic acid (NPA), suggesting that accumulation of newly synthesized auxin in these regions is due to an NPA-sensitive auxin transport loop (Fig. 2H, e). In *pin2* mutant background the induced responses are restricted to the lateral root cap and epidermis, confirming a role for basipetal transport of auxin towards provascular cells in the root meristem (Supplementary Fig. 3).

Our results suggest that the capacity to circulate auxin through loop-oriented PIN efflux facilitators regulates meristem size. Such a loop system can redeploy auxin and hence operate at least partially independently from shoot auxin supply, consistent with the ability of isolated root systems to maintain growth without external auxin application. In a wider perspective, these data support that PIN protein localization predicts auxin transport routes.

PIN genes regulate cell expansion and root elongation zone size

The capacity to stimulate cell expansion upon external application has been a defining property for auxins¹⁹. Recent analysis of tropic responses revealed that *PIN* genes play a part in differential auxin distribution, which is accompanied by differential cell expansion without cell division occurs in the elongation zone located proximal to the meristem zone (Fig. 2A).

Final cell size is affected in several *pin* mutants, but no additive effects occur in mutant combinations, suggesting that *PIN* gene action on cell expansion is complex. Nevertheless, in single allelic combinations of *pin1pin3* and *pin3pin7*, the change in final cell length is the major factor accounting for the reduction in root length (Supplementary Fig. 2, red fonts). In contrast to mature cell size, the size of cells immediately after departure from the meristem is not affected in any *pin* mutant combination (Supplementary Fig. 2).

We observed a reduction in the size of the elongation zone in *pin* mutant combinations, which mostly correlated well with reduction in meristem size. However, in *pin3pin4pin7* only the elongation zone size is reduced (Supplementary Fig. 2, blue fonts). These data

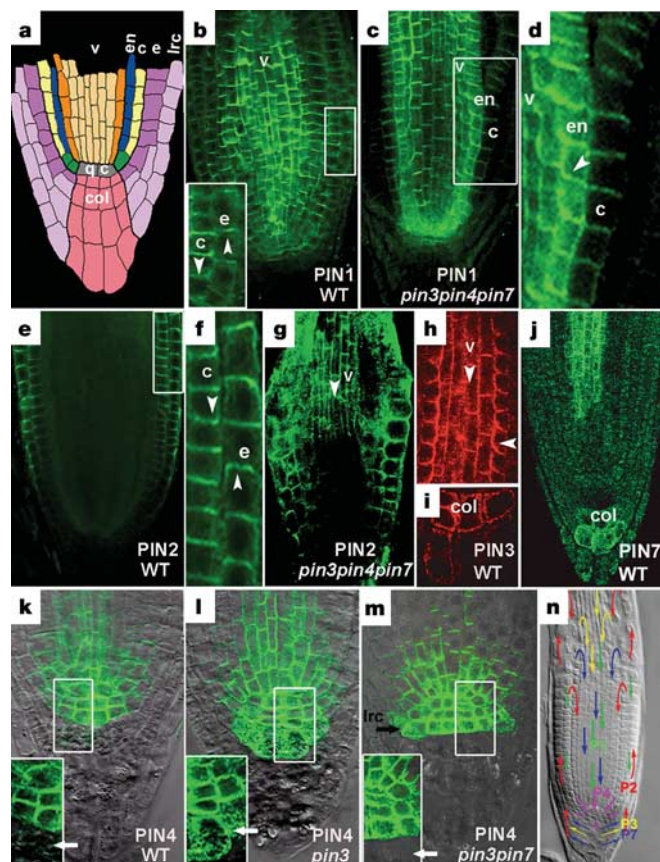


Figure 1 PIN expression and protein localization in roots of wild-type (WT) and *pin* mutant *Arabidopsis*. **a**, *Arabidopsis* root meristem with columella (col), quiescent centre (qc), lateral root cap (lrc), epidermis (e), cortex (c), endodermis (en) and vascular bundle (v). **b–d**, PIN1 immunolocalization: **b**, wild type; inset shows boxed area enlarged. Arrowheads depict polar localization. **c, d**, *pin3pin4pin7* (Allelic Combination 1 (AC1)), boxed area in **c** is enlarged in **d**. Arrowheads in **d**, PIN1 upregulation in the endodermis. **e–g**, PIN2 protein: **e, f**, Wild type; boxed area in **e** is enlarged in **f**. Arrowheads in **f**, apical PIN2 in the epidermis and basal in the cortex. **g**, *pin3pin4pin7* (AC1); arrowhead shows ectopic expression. **h, i**, PIN3 protein in wild-type root meristem. Arrowheads in **h**, PIN3 in the pericycle. **j**, PIN7–green fluorescent protein (GFP) fusion. **k–m**, PIN4 protein: **k**, wild type; white arrow, wild-type differentiated columella cell without PIN4. **l**, *pin3* (white arrow); **m**, *pin3pin7* (AC1): black arrow, PIN4 expansion to the lateral root cap; white arrow as in **k**. Insets: enlargements of boxed areas showing details of PIN4 localization. **n**, Localization of PIN proteins suggests auxin transport routes. PIN1 (green), PIN2 (red), PIN3 (yellow), PIN4 (violet) and PIN7 (blue). Immunolocalization signals are green in **b–g** and **k–m** and red in **h, i**.

indicate that the region where cell elongation occurs can be independently controlled by *PIN* gene activity.

Together, our data reveal that modulation of *PIN* activities can separately affect meristem size, elongation zone size and final cell size. These effects are not additive but probably result from interactions between changes in auxin distribution and transcriptional or translational responses influencing carrier components. We conclude that *PIN*-mediated modulation of auxin distribution controls both cell division and cell elongation and thereby contributes to the 'organizing' role of auxin in organ growth.

PIN genes regulate *PLT* expression and pattern the distal root meristem

Polar auxin transport is a major contributor to root meristem patterning in *Arabidopsis*, and the specification of distal cell types correlates well with the auxin response maximum¹. Close to the auxin maximum, the quiescent centre maintains surrounding cells as stem cells (Fig. 1a). Quiescent centre and stem cell specification require *SHR* and *SCR*, putative GRAS family transcription factors^{27,28} and the redundantly acting *PLT1* and *PLT2* AP2-domain putative transcription factors¹⁷. *PLT* transcript accumulation is

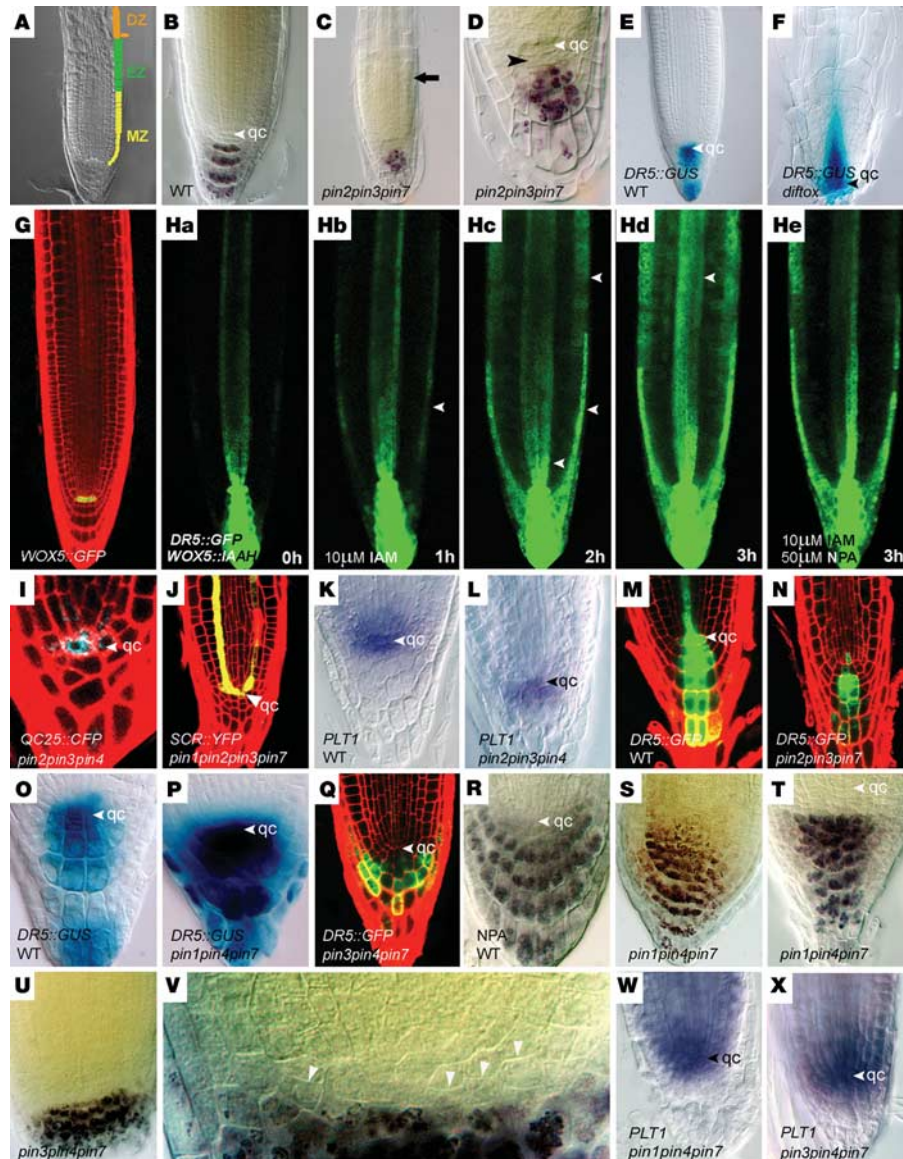


Figure 2 *PIN* genes control meristem size and patterning in *Arabidopsis* roots. **A–F**, Meristem size control. **A**, Wild-type meristem zone (MZ), elongation zone (EZ) and differentiation zone (DZ). **B**, Columella cell staining. **C, D**, *pin2pin3pin7*^(AC1): arrow in **C**, border of MZ and EZ; black arrowhead in **D**, columella stem cells. **E, F**, *DR5::GUS* expression in wild type (**E**) and *prCP1::DT-A^{tsM}* (root cap expressed diphtheria toxin, 'diftoxin') (**F**). **G**, *WOX5* promoter specificity for quiescent centre. **H, a–e**, *DR5::GFP* in *WOX5::IAAH* plants. **a**, Control. **b–d**, *DR5::GFP* upregulation (arrowheads) after IAM application. **e**, No provascular upregulation on 50 μ M NPA. **I–L**, Patterning in *pin2* mutant combinations. **I**, *QC25::CFP* (cyan fluorescent protein) in *pin2pin3pin4*^(AC2). **J**, *SCR::YFP* (yellow fluorescent protein) in *pin1pin2pin3pin7*^(AC1). **K, L**, In wild type (**K**) and in *pin2pin3pin4*^(AC2) (**L**), *PLT1* transcript is restricted to the quiescent centre and stem cells.

M, N, *DR5::GFP* in wild type (**M**) and *pin2pin3pin7*^(AC1) (**N**). **O–X**, *pin* mutants with patterning defects. Wild type treated with NPA (**R**). In 33% of *pin1pin4pin7*^(AC1) *DR5::GUS* (**P**), starch-granule-containing columella cells (**S**) and *PLT1* expression (**W**) expand proximally (white arrowhead). In *pin1pin4pin7*^(AC2), columella cells also expand laterally (**T**). In *pin3pin4pin7*^(AC1), mature columella cells (**U**), their stem cells (arrowheads in **V**), *DR5::GFP* expression (**Q**) and *PLT1* mRNA (**X**) expand laterally. *PLT1* transcript: whole-mount *in situ* hybridization (blue/purple signals). Differentiated columella cells: starch granule staining (purple). *DR5::GUS*, blue. GFP fluorescence, green. qc, quiescent centre. **A–F, K, L, O, P, R–X**, Nomarski optics; **G, H, I, J, M, N, Q**, CLSM after propidium iodide staining.

correlated with auxin accumulation in embryos and primary and lateral roots, and depends on auxin response factors¹⁷. The identification of these critical factors for specification of distal cell types as well as the highly organized PIN gene expression domains in the

root primordium set the stage for an analysis of the role of *PIN* genes in pattern formation.

In the previously discussed *pin2* mutant combinations, meristem size is affected but distal patterning is normal, as judged by the presence of columella stem cells (Fig. 2D), the QC25 marker for the quiescent centre (Fig. 2I), *SCR* promoter activity (Fig. 2J), *PLT1* (ref. 17) transcript distribution (Fig. 2K, L) and the DR5 auxin response marker (Fig. 2M, N).

In *pin1pin4pin7*, however, the auxin response maximum and the starch granules that mark the differentiated columella cells shift proximally in 40% of roots (Fig. 2P, S). In *pin3pin4pin7*, the auxin response maximum shifts laterally (Fig. 2Q), associated with inappropriate lateral up-regulation of PIN1 (Fig. 1c, d). Correlated with this lateral expansion, a broadening of the columella domain including its stem cells occurs (Fig. 2U, V). Consistent with changes in distal patterning, *PLT1* messenger RNA shifts proximally in *pin1pin4pin7* (Fig. 2W) and expands laterally in *pin3pin4pin7* (Fig. 2X).

The phenotypes of both triple mutants resemble effects of treatment with inhibitors of polar auxin transport (Fig. 2R) and support a role for PIN proteins in focusing and stabilizing an auxin maximum in both proximo-distal and lateral dimensions. Our data strongly suggest that the PIN proteins in this way focus expression of the auxin-inducible *PLT* genes in the distal root region, which specifies the position of the quiescent centre and stem cells.

PIN genes restrict *PLT* mRNA and root identity to the basal embryo pole

We addressed whether *PIN* genes control *PLT* expression and root stem cell patterning during embryogenesis. Distal quiescent centre and columella cells originate from daughter cells of the hypophysis at the early globular stage of embryogenesis (Fig. 3a–d). At that stage an auxin perception maximum is detected in the hypophysis⁴.

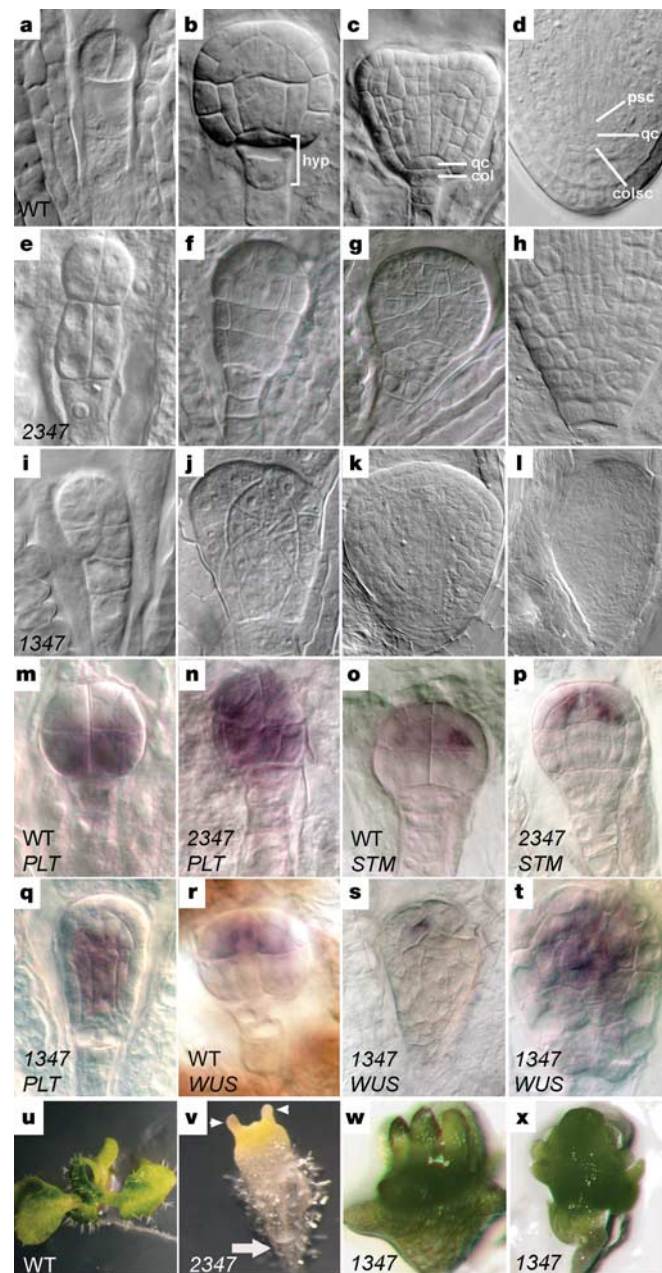


Figure 3 *PIN* genes and embryonic patterning. **a–d**, Wild-type embryo at 2-cell (**a**) globular (**b**) early heart (**c**) and torpedo stage (**d**). hyp, hypophysis; qc, quiescent centre; col, columella; sc, stem cells; psc, provascular stem cells. **e–h**, *pin2pin3pin4pin7*(*ACT1*) embryos: **e**, aberrant divisions in basal cells at octant stage; **f**, basal cell duplication at globular stage; **g**, **h**, quiescent centre and stem cell division defects at heart (**g**) and torpedo stages (**h**). **i–l**, *pin1pin3pin4pin7*(*ACT1*) embryos with abnormal basal (**i**) and apical cell divisions (**j–l**). **m–t**, Gene expression in *pin* quadruples: **m**, *PLT1* mRNA in wild-type preglobular embryos is restricted to basal cells; **n**, ubiquitous *PLT1* in *pin2pin3pin4pin7*(*ACT1*) proembryos; **p**, **o**, *STM* in *pin2pin3pin4pin7*(*ACT1*) (**p**) and wild type (**o**); **q**, *PLT1* in *pin1pin3pin4pin7*(*ACT1*); **r–t**, In comparison to wild type (**r**), *WUS* is slightly reduced (**s**) or expands (**t**). **u–x**, Explanted embryos: **u**, wild type develops normal seedlings; **v**, *pin2pin3pin4pin7*(*ACT1*) develops reduced cotyledons (arrowheads) and ectopic root hairs (arrow); **w**, **x**, *pin1pin3pin4pin7*(*ACT1*) explants have reduced root development and ectopic shoot-like structures at the apex (green tissues). *PLT1*, *STM* and *WUS* transcripts: *in situ* hybridization (purple). Images: Nomarski optics.

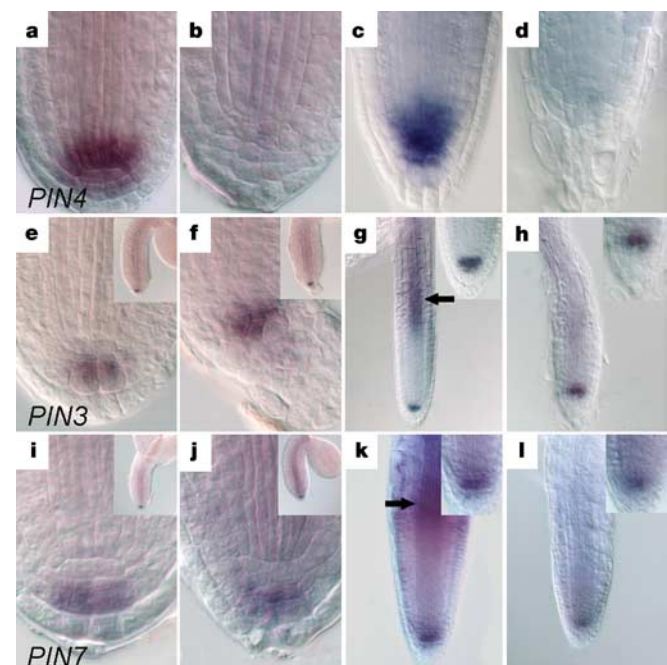


Figure 4 *PLT* genes regulate *PIN* transcript levels. **a–d**, *PIN4* mRNA localization. *PIN4* transcript in the quiescent centre, stem cells and provascular initials of wild-type embryo (**a**) and seedlings (**c**). No detectable transcripts in *pld1pld2* double-mutant embryos (**b**) and seedlings (**d**). **e–l**, *PIN3* and *PIN7* mRNA in columella cells and provascular region of wild-type embryo (**e**, **i**) and seedlings (**g**, **k**, arrows). *pld1pld2* with normal levels in embryonic columella (**f**, **j**) and seedlings (**h**, **l**). Reduction of provascular expression in seedlings (**h**, **l**). Images: Nomarski optics.

Clonal analysis and stereotyped cell division patterns show that the proximal stem cells are recruited around mid-heart stage (Fig. 3c)²⁹.

We found no penetrant embryonic defects in single, double and triple *pin* mutants and in *pin1pin2pin3pin7* and *pin1pin2pin3pin4*. However, in lines segregating *pin2pin3pin4pin7* and *pin1pin3pin4pin7* mutants, marked changes were observed in embryonic division patterns and gene expression (Fig. 3e–t and Supplementary Table 2). *pin2pin3pin4pin7* mutant embryos exhibit cell division defects mainly in the basal region (Fig. 3e–h) and ~30% produce viable seedlings. In contrast, *pin1pin3pin4pin7* mutants are embryo lethal and cell division defects occur in apical and basal embryo regions⁴ (Fig. 3i–l). The significance of the altered cell divisions in both quadruple-mutant embryos in terms of patterning was investigated by locating the expression domain of marker genes for root and shoot identity. *PLT1* transcript marks the basal domain of the octant stage embryo (Fig. 3m) and restricts to the quiescent centre and stem cell domain before mid-heart stage¹⁷. *STM* and *WUS* are required for shoot meristem function and are transcribed in shoot meristem precursor cells from early embryogenesis onwards^{30,31}.

pin2pin3pin4pin7 mutants contain both aberrant cell divisions and high levels of *PLT1* transcript throughout the embryo from the 16-cell stage onwards (Fig. 3n). In contrast, *PLT1* mRNA is correctly excluded from the apical region in *pin1pin3pin4pin7* mutants (Fig. 3q). Interestingly, the mRNA localization of *WUS* and *STM* does not change in *pin2pin3pin4pin7* mutants (Fig. 3o, p), whereas in *pin1pin3pin4pin7*, *WUS* transcripts are either slightly reduced (Fig. 3s) or expanded in apical embryonic cells (Fig. 3t).

In explanted wild-type embryos, the shoot and root apical meristems develop normally (Fig. 3u). In ~40% of *pin2pin3pin4pin7* mutants, the explants develop reduced cotyledons and root hairs emerge at more apical positions (Fig. 3v and Supplementary Table 2). In ~80% of *pin1pin3pin4pin7* mutants, the explants completely arrest root growth and expand the shoot domain (Fig. 3w, x and Supplementary Table 2). The explant phenotypes of both mutants match the observed expansion of *PLT1* and *WUS* domains, because ectopic expression of these genes promotes root and shoot identity, respectively^{17,32}. *PLT* gene expression is strongly dependent on the joint action of PIN proteins, in line with its dependence on auxin response factors and the correlation between *PLT* expression and auxin accumulation¹⁷. Most probably the *PLT* expression domain is regulated by PIN4 and PIN7, which are

appropriately positioned for basal auxin transport at the pre-globular stage⁴. PIN3 and PIN2 are not expressed at this stage but ectopic PIN2 mRNA can be detected in *pin3pin4pin7* siliques (data not shown), suggesting that ectopic expression of PIN members in embryos provides a remarkably versatile compensatory mechanism for the loss of PIN4 and PIN7.

PIN-regulated early *PLT*, *WUS* and *STM* transcription suggests that auxin transport in the embryo regulates the proper expression of critical root and shoot stem cell regulators. The notion that early cellular asymmetries in membrane localization of PIN proteins are translated into the patterning of embryonic stem cell domains via regulation of auxin flux provides a conceptual framework for initial events in plant embryogenesis.

PLT genes regulate PIN gene expression in the root meristem

The *PLT* genes are required for specification of the stem cell niche and convey root identity when ectopically expressed¹⁷. Therefore we asked whether the *PLT* proteins could regulate the root-specific distribution of *PIN* transcripts, thereby fine-tuning the position of the stem-cell-associated auxin maximum, and cell division and cell expansion domains. Remarkably, *PIN4* transcript, which overlaps with the *PLT* transcripts in wild type, is undetectable in 94% and aberrant in 6% of the embryos and seedlings of *plt1plt2* (Fig. 4a–d). *PIN3* and *PIN7* transcripts are normal in columella cells (Fig. 4e–l) but are markedly reduced in the provascular domain of the post-embryonic root elongation zone (Fig. 4g, h, k, l). Thus, the *PLT* genes control *PIN* mRNA distribution.

Discussion

Our findings suggest an elegant mechanism for embryonic root primordium formation and stabilization. In our model, PIN proteins restrict *PLT* expression in the basal embryo region to initiate root primordium formation (Fig. 5a). In turn, *PLT* genes maintain *PIN* transcription, which stabilizes the position of the distal stem cell niche (Fig. 5b). At a distance from the auxin maximum, *PLT* genes maintain PIN3 and PIN7, which reinforce provascular acropetal auxin flux. In this way a 'reflux' loop is created that controls auxin distribution in the growing primordium and meristem (Fig. 5b, c). The loop stabilizes the auxin maximum and the *PLT*-dependent stem cell domain in the distal root tip. Moreover, it localizes meristem and cell expansion zones in the proximal meristem and regulates final cell size (Fig. 5c).

Reporter genes and direct auxin measurements are consistent with the presence of a transport-regulated auxin gradient in the root meristem¹, but local auxin biosynthesis and catabolism may also contribute to the auxin concentration profile³³. Furthermore, auxin response may be regulated differently in the different root zones by processes such as differential SCF^{TIR1}-mediated proteolysis³⁴. □

Methods

Materials

All *pin* mutants were in Columbia (Col-0) background, except for the Enkheim allele *pin1-1*. *pin1-1*, *pin1En134*; *eir1-1*, *pin2En701*; and *pin3-3*, *pin4-2* were described in refs 7–10. *pin3 salk_005544* and *pin7 salk_048791* were provided by the Signal Insertion Mutant Library (<http://signal.salk.edu/cgi-bin/tdnaexpress/>). *plt1-4* and *plt2-2* alleles were described in ref. 17.

Triple mutants were generated by crossing double mutants sharing one allele, while quadruple mutants were generated by crossing triple homozygotes with two alleles in common. *DR5::GUS* (*DR5*-β-glucuronidase) was described in ref. 1 and was crossed with *pin1*, *pin2*, *pin4* and *pin7*. Homozygous lines were then used to generate double and triple mutants homozygous for *DR5::GUS*. *DR5-GFP*, described in ref. 26, was transformed to wild type, *pin2pin3pin7* and *pin3pin4pin7*. Promoter constructs of *QC25::CFP* and *SCR::YFP* were transformed to *pin2pin3pin4* and to *pin1pin2pin3pin7* respectively. *WOX5::GFP* and *WOX5::IAAH* constructs were generated by fusing a 4.5-kb *WOX5* (ref. 24) promoter fragment in front of *GFP* or *IAAH*²⁵ in the pGreenII0229 (ref. 35) vector and transformed into the wild type or *DR5::GFP* lines.

Phenotype analysis and microscopy

Plant material for light microscopy was prepared as in ref. 36. Starch granules and β-glucuronidase activity were visualized as in ref. 36. For embryo phenotype analysis,

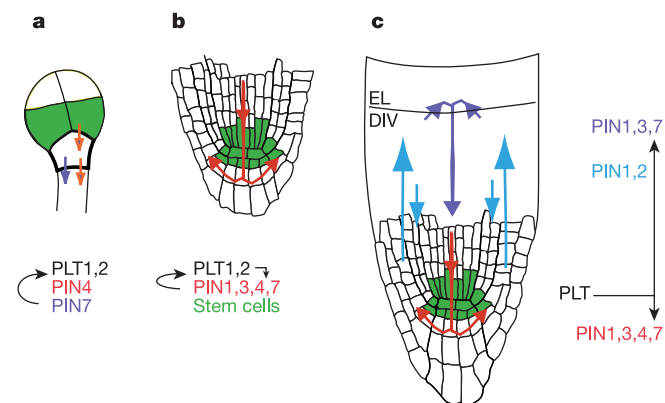


Figure 5 Model for primordium formation by PIN–PLT interactions. **a**, PIN-mediated root primordium specification by restriction of *PLT* transcripts in octant/16-cell embryo stage. **b**, At later stages of embryogenesis, PIN action further restricts *PLT* transcripts to define the stem cell region and *PLT* genes start controlling root-specific *PIN* gene expression. **c**, In post-embryonic roots, PIN-mediated auxin transport stabilizes the stem cell region and regulates cell division (DIV) in the meristem zone and cell expansion in the elongation zone (EL). *PLT* genes control several members of the *PIN* gene family to generate primordium-specific auxin distribution.

ovules were dissected and embryos were prepared as described in ref. 29. Embryo explants were cultured on petri dishes containing one-half GM medium and grown on the same medium for 4 weeks. For exogenous auxin application, wild type and multiple mutants were grown for 4 days (d) and transferred to medium supplemented with 100–200 nM NAA for another 3 d. Induction of auxin biosynthesis was performed by transferring plants to medium containing 10 μ M of the precursor IAM for 0–3 h. Images were taken on a Zeiss Axioskop with a Nikon DXM1200 digital camera.

In situ hybridization and immunolocalization

For whole-mount *in situ* hybridization⁸, gene-specific 600-base-pair complementary DNA fragments for *PIN4*, *PIN3* and *PIN7* were used as probes. *PLT1* probe was described in ref. 17 and *STM* and *WUS* probes were synthesized using the complete cDNA sequence. Immunolocalizations were performed on whole-mount roots of 5–7-day-old seedlings^{7,37}. PINs antibodies and reagents were described in refs 5–8. Imaging was performed using a Leica SP2 inverted confocal microscope and accompanying software.

Received 5 July; accepted 5 November 2004; doi:10.1038/nature03184.

1. Sabatini, S. *et al.* An auxin-dependent distal organizer of pattern and polarity in the *Arabidopsis* root. *Cell* **99**, 463–472 (1999).
2. Benková, E. *et al.* Local, efflux-dependent auxin gradients as a common module for plant organ formation. *Cell* **115**, 591–602 (2003).
3. Reinhardt, D. *et al.* Regulation of phyllotaxis by polar auxin transport. *Nature* **426**, 255–260 (2003).
4. Friml, J. *et al.* Efflux-dependent auxin gradients establish the apical-basal axis of *Arabidopsis*. *Nature* **426**, 147–153 (2003).
5. Gälweiler, L. *et al.* Regulation of polar auxin transport by AtPIN1 in *Arabidopsis* vascular tissue. *Science* **282**, 2226–2230 (1998).
6. Müller, A. *et al.* AtPIN2 defines a locus of *Arabidopsis* for root gravitropism control. *EMBO J.* **17**, 6903–6911 (1998).
7. Friml, J., Wisniewska, J., Benková, E., Mendgen, K. & Palme, K. Lateral relocation of auxin efflux regulator PIN3 mediates tropism in *Arabidopsis*. *Nature* **415**, 806–809 (2002).
8. Friml, J. *et al.* AtPIN4 mediates sink-driven auxin gradients and root patterning in *Arabidopsis*. *Cell* **108**, 661–673 (2002).
9. Luschig, C., Gaxiola, R. A., Grisa, P. & Fink, G. R. EIR1, a root-specific protein involved in auxin transport, is required for gravitropism in *Arabidopsis thaliana*. *Genes Dev.* **12**, 2175–2187 (1998).
10. Utsuno, K., Shikanai, T., Yamada, Y. & Hashimoto, T. Agr, an agravitropic locus of *Arabidopsis thaliana*, encodes a novel membrane-protein family member. *Plant Cell Physiol.* **39**, 1111–1118 (1998).
11. Chen, R. *et al.* The *Arabidopsis thaliana* AGRVITROPIC 1 gene encodes a component of the polar-auxin-transport efflux carrier. *Proc. Natl Acad. Sci. USA* **95**, 15112–15117 (1998).
12. Noh, B., Murphy, A. S. & Spalding, E. P. *Multidrug Resistance*-like genes of *Arabidopsis* required for auxin transport and auxin-mediated development. *Plant Cell* **13**, 2441–2454 (2001).
13. Noh, B., Bandyopadhyay, A., Peer, W. A., Spalding, E. P. & Murphy, A. S. Enhanced gravi- and phototropism in plant *mdr* mutants mislocalizing the auxin efflux protein PIN1. *Nature* **423**, 999–1002 (2003).
14. Steinmann, T. *et al.* Coordinated polar localization of auxin efflux carrier PIN1 by GNOM ARF GEF. *Science* **286**, 316–318 (1999).
15. Geldner, N., Friml, J., Stierhof, Y. D., Jürgens, G. & Palme, K. Auxin transport inhibitors block PIN1 cycling and vesicle trafficking. *Nature* **413**, 425–428 (2001).
16. Geldner, N. *et al.* Partial loss-of-function alleles reveal a role for GNOM in auxin transport-related, post-embryonic development of *Arabidopsis*. *Development* **131**, 389–400 (2004).
17. Aida, M. *et al.* The *PLETHORA* genes mediate patterning of the *Arabidopsis* root stem cell niche. *Cell* **119**, 109–120 (2004).
18. Sieberer, T. *et al.* Post-transcriptional control of the *Arabidopsis* auxin efflux carrier EIR1 requires AXR1. *Curr. Biol.* **10**, 1595–1598 (2000).

19. Srivastava, L. *Plant Growth and Development: Hormones and Environment* (Academic, New York, 2002).
20. Okada, K., Ueda, J., Komaki, M., Bell, C. & Shimura, Y. Requirement of the auxin polar transport system in early stages of *Arabidopsis* floral bud formation. *Plant Cell* **3**, 677–684 (1991).
21. Davies, P. J., Doro, J. A. & Tarbox, A. W. The movement and physiological effect of indoleacetic acid following point applications to root tips of *Zea mays*. *Physiol. Plant.* **36**, 333–337 (1976).
22. Rashotte, A. M., Brady, S. R., Reed, R. C., Ante, S. J. & Muday, G. K. Basipetal auxin transport is required for gravitropism in roots of *Arabidopsis*. *Plant Physiol.* **122**, 481–490 (2000).
23. Tsugeki, R. & Fedoroff, N. V. Genetic ablation of root cap cells in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **96**, 12941–12946 (1999).
24. Haecker, A. *et al.* Expression dynamics of *WOX* genes mark cell fate decisions during early embryonic patterning in *Arabidopsis thaliana*. *Development* **131**, 657–668 (2004).
25. Kares, C., Prinsen, E., Van Onckelen, H. & Otten, L. IAA synthesis and root induction with *iaa* genes under heat shock promoter control. *Plant Mol. Biol.* **15**, 225–236 (1990).
26. Ottenschläger, I. *et al.* Gravity-regulated differential auxin transport from columella to lateral root cap cells. *Proc. Natl Acad. Sci. USA* **100**, 2987–2991 (2003).
27. DiLaurenzio, L. *et al.* The SCARECROW gene regulates an asymmetric cell division that is essential for generating the radial organization of the *Arabidopsis* root. *Cell* **86**, 423–433 (1996).
28. Helariutta, Y. *et al.* The *SHORT-ROOT* gene controls radial patterning of the *Arabidopsis* root through radial signaling. *Cell* **101**, 555–567 (2000).
29. Scheres, B. *et al.* Embryonic origin of the *Arabidopsis* primary root and root meristem initials. *Development* **120**, 2475–2487 (1994).
30. Mayer, K. F. *et al.* Role of WUSCHEL in regulating stem cell fate in the *Arabidopsis* shoot meristem. *Cell* **95**, 805–815 (1998).
31. Long, J. A. & Barton, M. K. The development of apical embryonic pattern in *Arabidopsis*. *Development* **125**, 3027–3035 (1998).
32. Gallois, J. L., Nora, F. R., Mizukami, Y. & Sablowski, R. WUSCHEL induces shoot stem cell activity and developmental plasticity in the root meristem. *Genes Dev.* **18**, 375–380 (2004).
33. Ljung, K. *et al.* Biosynthesis, conjugation, catabolism and homeostasis of indole-3-acetic acid in *Arabidopsis thaliana*. *Plant Mol. Biol.* **50**, 309–332 (2002).
34. Estelle, M. Auxin signaling and regulated protein degradation. *Trends Plant Sci.* **9**, 302–308 (2004).
35. Hellens, R. P., Edwards, E. A., Leyland, N. R., Bean, S. & Mullineaux, P. M. pGreen: a versatile and flexible binary Ti vector for Agrobacterium-mediated plant transformation. *Plant Mol. Biol.* **42**, 819–832 (2000).
36. Willemsen, V., Wolkenfelt, H., de Vrieze, G., Weisbeek, P. & Scheres, B. The *HOBBIT* gene is required for formation of the root meristem in the *Arabidopsis* embryo. *Development* **125**, 521–531 (1998).
37. Friml, J., Benkova, E., Mayer, U., Palme, K. & Muster, G. Automated whole-mount localization techniques for plant seedlings. *Plant J.* **34**, 115–124 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank N. Fedoroff and R. Tsugeki for seeds of *pRCPI::DT-A^{EM}*; P. Hogeweg, S. Smekens and members of the root laboratory for critical reading of the manuscript; and F. Kindt and R. Leito for artwork. K.P. acknowledges R. Nitschke and the Life Imaging Center (University of Freiburg). This work was supported by an NWO-PIONIER grant (I.B., J.X., M.W. and B.S.) and by the Deutsche Forschungsgemeinschaft and the FCI (I.P. and K.P.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.S. (b.scheres@bio.uu.nl).

The heating of gas in a galaxy cluster by X-ray cavities and large-scale shock fronts

B. R. McNamara¹, P. E. J. Nulsen^{2,*}, M. W. Wise³, D. A. Rafferty¹, C. Carilli⁴, C. L. Sarazin⁵ & E. L. Blanton^{5,6}

¹Astrophysical Institute and Department of Physics & Astronomy, Ohio University, Clipping Laboratory, Athens, Ohio 45701, USA

²Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA

³MIT Center for Space Research, 1 Hampshire Street, Cambridge, Massachusetts 02139, USA

⁴National Radio Astronomy Observatory, Very Large Array, Socorro, New Mexico 87801, USA

⁵Astronomy Department, University of Virginia, Box 3818, Charlottesville, Virginia 22903, USA

⁶Institute for Astrophysical Research, Boston University, Boston, Massachusetts 02215, USA

* Present address: University of Wollongong, Wollongong, New South Wales, Australia

Most of the baryons in galaxy clusters reside between the galaxies in a hot, tenuous gas¹. The densest gas in their centres should cool and accrete onto giant central galaxies at rates of 10–1,000 solar masses per year¹. No viable repository for this gas, such as clouds or new stars, has been found¹. New X-ray observations, however, have revealed far less cooling below X-ray temperatures than expected², altering the previously accepted picture of cooling flows. As a result, most of the gas must be heated to and maintained at temperatures above ~ 2 keV (ref. 3). The most promising heating mechanism is powerful radio jets emanating from supermassive black holes in the central galaxies of clusters⁴. Here we report the discovery of giant cavities and shock fronts in a distant ($z = 0.22$) cluster caused by an interaction between a

radio source and the hot gas surrounding it. The energy involved is $\sim 6 \times 10^{61}$ erg, the most powerful radio outburst known. This is enough energy to quench a cooling flow for several Gyr, and to provide $\sim 1/3$ keV per particle of heat to the surrounding cluster.

Cavities with diameters ranging from a few to a few tens of kiloparsecs have been found in the hot gas surrounding nearly two dozen galaxies, groups and clusters⁵. Their enthalpy (free energy) lies between 10^{55} – 10^{60} erg and scales in proportion to the cooling X-ray luminosity and the radio power of the host system⁵. The cavities and weak shocks in half of these systems are currently injecting enough energy into the hot gas to balance radiation losses (cooling)^{6,7}, but it is unclear whether they can quench cooling over longer timescales. Systems without cavities today could have been heated in the past by powerful but relatively rare outbursts⁸. However, before the discoveries of large-scale cavities and shocks discussed here and elsewhere⁹, little evidence existed to support this conjecture.

Giant cavities, each roughly 200 kpc in diameter, were found in a Chandra X-ray image of the optically poor cluster MS0735.6+7421 (Fig. 1). The centre of the cluster harbours a cD galaxy that hosts a radio source roughly 550 kpc in size. The radio lobes fill the cavities, which suggests, as in other clusters, that the gas was displaced and compressed by the advancing radio source. Both the cavities and the radio source dwarf the cD galaxy, which is itself a member of the class of the largest galaxies in the Universe.

The average pressure surrounding the cavities (Fig. 2) is $p \approx 6 \times 10^{-11}$ erg cm⁻³. The work required to inflate each cavity against this pressure is $pV \approx 10^{61}$ erg, where V is the volume of the cavity. The enthalpy of the cavities can approach $4pV$ per cavity, depending on the equation of state of the gas filling them, giving a value of about 8×10^{61} erg. To place this figure in perspective, it exceeds the enthalpy in the Perseus cluster⁶, Cygnus A⁵, and M87 (ref. 7) by roughly 250 times, 15 times, and more than four orders of magnitude, respectively. Only the energy of the recently discovered shock front in Hydra A⁹ falls within an order of magnitude of this value.

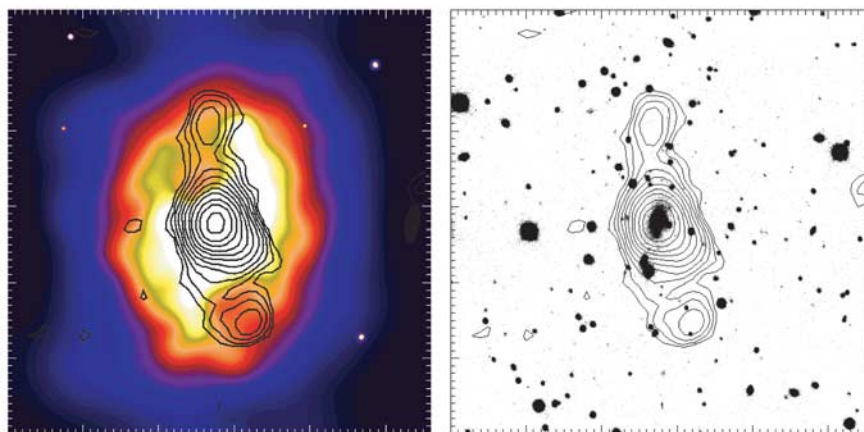


Figure 1 The relationships between the X-ray, radio and optical emissions of the cluster. The smoothed X-ray image (left) and optical image (right) are superposed with the 1.4-GHz radio contours. The 40-ks X-ray image was obtained with the Chandra X-ray Observatory on 1 December 2003. Approximately 75,000 useful X-ray photons were detected. The X-ray surface brightness depressions (cavities) are between 10%–20% fainter than the surrounding X-ray emission to the north and south of the centre. Most of the cluster's emission emerges from an elliptical structure bounded by a shock front. We assume a flat cosmology, with $H_0 = 70$ km s⁻¹ Mpc⁻¹ and $\Omega_m = 0.3$, corresponding to a ratio of linear to angular size of 3.5 kpc arcsec⁻¹ at the redshift of the cluster throughout this paper. The ≈ 4 -arcsec resolution radio map was made with the Very

Large Array telescope in the C configuration. The cavities are filled with radio emission. Assuming spherical cavities whose edges lie at the edges of the rims, each is roughly an arcminute in diameter (200 kpc) centred approximately 125 kpc to the northeast and to the southwest of the cluster centre: right ascension, RA = 07 h 41 min 44.0 s, declination, Dec. = $+74^\circ 14' 38.3''$ (J2000). The radio contour levels are $2 \times 10^{-4} \times (-1, 1, 1.4, 2, 2.8, 4, 5.7, 8, 11, 16, 22, 32)$ Jy per beam. The radio-contour levels on the visual image are $2 \times 10^{-4} \times (-1, 1, 1.4, 2, 2.8, 4, 5.7, 8, 11, 16, 22)$ Jy per beam. The R -band surface brightness of the cD galaxy peaks at its centre at roughly 19.5 mag arcsec⁻² and diminishes with radius in $r^{1/4}$ -law fashion, reaching 25.5 mag at 105 kpc. Each image is 250×250 arcsec (875×875 kpc) on a side.

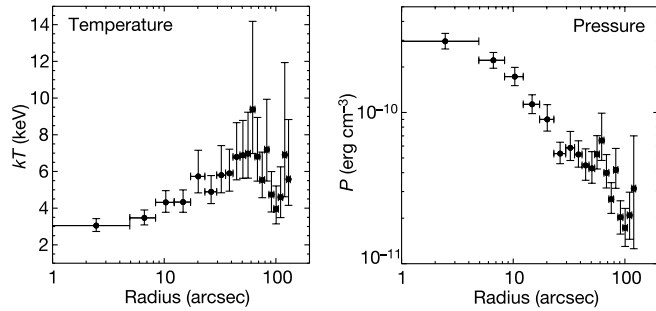


Figure 2 Projected temperature and pressure profiles of the cluster gas. The vertical error bars are 90% confidence intervals, and the horizontal bars represent the bin sizes. The spectra used to construct these profiles were extracted from the X-ray image in circular apertures centred on the cD galaxy. The spectra were modelled as thermal emission from a single-temperature plasma of uniform metallicity, attenuated by the foreground column of neutral hydrogen in our Galaxy. The average metallicity of the gas was found to be 0.4 times the solar value. The temperature and pressure profiles are complex. The coolest gas located in the centre of the cluster is roughly 3 keV. The temperature rises with increasing radius, reaching an average of about 7 keV between 50 and 80 arcsec. Beyond 80 arcsec the temperature drops to roughly 5 keV, although the level of this drop is uncertain. The pressure is highest in the centre, $3 \times 10^{-10} \text{ erg cm}^{-3}$, and declines smoothly with radius in roughly power-law fashion until reaching a radius of roughly 70 arcsec. The pressure there rises abruptly at the shock front.

The bright elliptical region surrounding the X-ray cavities strongly resembles the hot cocoon of jet-powered radio-source models^{10,11}. In this interpretation, the relatively sharp edge of the cocoon lies at the location of an enveloping shock (Fig. 3). The two red ‘hot spots’ in Fig. 4 show that the gas near the cavities is being heated by the shocks. In contrast, the gas surrounding the cavities in other systems such as Hydra A⁴ and Perseus⁶ is relatively cool, suggesting that buoyancy, not excess pressure, is driving their outward advance.

The shock properties were determined using a spherical hydrodynamic model of a point explosion at the centre of an initially isothermal, hydrostatic atmosphere (Fig. 3). The age and driving energy of the shock are $t_s \approx 1.04 \times 10^8 \text{ yr}$ and $E_s \approx 5.7 \times 10^{61} \text{ erg}$, respectively. The energy is proportional to the preshock temperature, which probably exceeds 5 keV. The spherical model underestimates the shocked volume, thus tending to underestimate total energy. Furthermore, because the cavities occupy a large fraction of the cocoon, they appear to be driving the shock, which undermines to some degree our assumption of a point explosion. Nevertheless, we expect the energy of the outburst to be within a factor of about two of the model estimate, and more likely to exceed it.

The shock energy is reassuringly close to the cavity enthalpy, and we adopt the shock energy as the probable value. The average jet power of the outburst is then $P_s = E_s/t_s \approx 1.7 \times 10^{46} \text{ erg s}^{-1}$, comparable to a powerful quasar radiating at the Eddington limit of a $\sim 2 \times 10^8 M_\odot$ black hole, where $M_\odot$ is the mass of the Sun. The flux density of the radio source at 1.4 GHz is 21 mJy (ref. 12), corresponding to a monochromatic luminosity of $4 \times 10^{40} \text{ erg s}^{-1}$, several orders of magnitude fainter than powerful quasars¹³. The ratio of average jet power to monochromatic radio power is $\sim 10^5$, enormously larger than the generally accepted factor of 10 to 100 (refs 13, 14) for powerful radio sources. Evidently, even relatively faint radio sources can be mechanically powerful⁵. Bright radiation from an active quasar is surprisingly absent both in the optical and X-ray bands. However, the cD galaxy harbours a $\sim 10^{42} \text{ erg s}^{-1}$ optical emission nebula extending over its inner 20 kpc (ref. 15), a type of nebula commonly found in cooling flows.

This outburst released enough energy to quench a $200 M_\odot \text{ yr}^{-1}$ cooling flow for several gigayears (Gyr), assuming all of the energy is

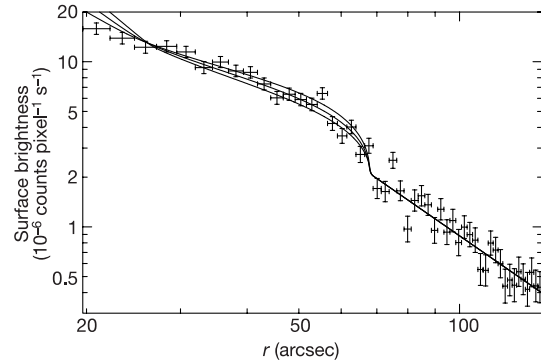


Figure 3 Projected 0.5–7.5-keV radial surface brightness profile of the cocoon region compared to shock-model predictions. The profile was measured in 30° sectors to the east and west, along the minor axis of the ellipse (position angles 90°–120° and 270°–300°), where the shock is relatively uniform. The feature at a radius of 69 arcsec (240 kpc) is consistent with being a weak shock. The surface brightness beyond is well fitted by the power law, $r^{-\beta}$ with $\beta = 2.24 \pm 0.29$ (90%). To be consistent with the surface brightness profile beyond the shock, the gas density initially has $\rho(r) \propto r^{-\eta}$, with $\eta = 1.62$, and the gravitational field was chosen to make the undisturbed atmosphere hydrostatic. The model assumes that the temperature of the unshocked gas is 5 keV. The Chandra 0.5–7.5-keV response was computed using XSPEC and an absorbed ‘mekal’ model with foreground column density $3.49 \times 10^{20} \text{ cm}^{-2}$, redshift 0.216 and abundance 0.4 times that of the Sun (results are insensitive to these parameters). The model surface brightness profiles are scaled to match the observed profile in the unshocked region. The three lines represent model profiles for shock Mach numbers of 1.41 and 1.41 ± 0.07 (increasing Mach number from bottom to top). The greatest source of uncertainty in the age is the preshock temperature, because the Mach number is not strongly model-dependent (in particular, if the cocoon is axially symmetric, the age estimate is not sensitive to projection effects). The temperature increase for a shock of this magnitude is expected to be roughly 30% above the preshock value. This jump is consistent with the data, but the uncertainty in the pre- and post-shock temperatures are too large to further constrain the shock properties. The vertical error bars are 90% confidence intervals, and the horizontal bars represent the bin sizes.

deposited within the ~ 50 -kpc cooling region of the cluster. The central cooling time of the gas is roughly 1 Gyr, so the cooling that ensues could establish a feedback cycle of heating and cooling, driven by accretion onto the central black hole^{8,16,17}. A similar process occurring in other clusters would in principle maintain the observed levels of hot gas with short cooling times, molecular gas¹⁸, and star formation¹⁹ in cD galaxies, while preventing the development of a more massive, steady cooling flow^{1,20}. The existence of bright nebular emission located in the cD galaxy is consistent with this picture¹⁵.

Much of the energy is, however, leaving the cooling region, bound for the cluster’s outskirts. The gas mass $5.5 \pm 0.7 \times 10^{13} M_\odot$ within 1 Mpc is being heated at the level of about 1/3 keV per particle. This one event alone then provides a substantial fraction of the 1–3 keV per particle of heating required to raise the entropy above the level of gravity alone (preheating)²¹. For a bolometric, unabsorbed X-ray luminosity of $1.1 \times 10^{45} \text{ erg s}^{-1}$ and a mean temperature of 4.5 keV, the cluster departs upward in luminosity by several times from the relatively tight correlation between X-ray luminosity and gas temperature^{22,20}. The shock power exceeds the cluster’s X-ray luminosity by 15 times, so the shock is surely capable of causing this departure. The time required to radiate away the shock energy of more than $E_s/L_x \approx 2 \text{ Gyr}$ is a substantial fraction of the age of the cluster. Therefore, this outburst will leave a persistent mark on the temperature and luminosity of the cluster, long after the cavities have disappeared. Assuming this event is not unique, substantial heating must have occurred recently in clusters, not just during an early preheating epoch²¹. Events of this nature would complicate the

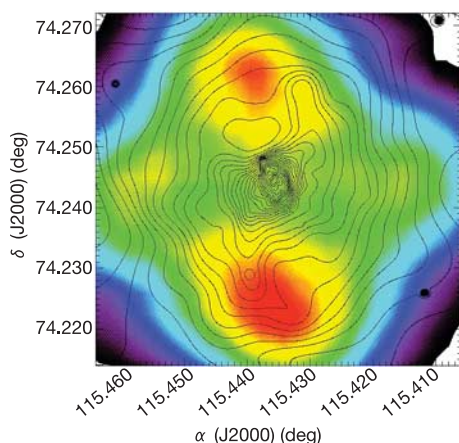


Figure 4 Temperature map of the central 200×200 arcsec of the cluster. Redder colours indicate hotter temperatures. The logarithmically spaced contours, ranging in surface brightness from $1\sigma = 1.0^{-4}$ counts $s^{-1} \text{ cm}^{-2} \text{ pixel}^{-2}$ to 100σ , show the locations of the cavities. The hottest regions are at the tips of the cavities, where the shocks are strongest.

use of X-ray temperature and luminosity functions to probe the large-scale structure and cosmology²².

The huge proportion of this event suggests that it was powered by accretion onto a black hole. Equating the shock energy to $0.1 Mc^2$ gives an estimate of the minimum accreted mass required to power the burst of $M \approx 3 \times 10^8 M_\odot$, itself the mass of a supermassive black hole. The relationship between galactic-bulge luminosity and black-hole mass²³ predicts that a $\sim 10^9 M_\odot$ black hole resides there (assuming the cD galaxy's absolute visual magnitude within a 35-kpc diameter is -22.4 ; ref. 24). The central black hole evidently accreted a substantial fraction of its own mass in only 10^8 yr, a remarkable growth rate for such a large black hole. Although a similar line of reasoning would apply to quasars, their ages and average jet powers have not been measured directly. Such a rapid rate of growth may be difficult to reconcile with the small scatter in the relation of the black-hole mass to the bulge mass²⁵, and is at variance with the view that the most massive black holes have evolved slowly in the recent past²⁶.

Finally, the magnetic field strengths in clusters are typically a few microgauss²⁷, and evidence is growing for the existence of large-scale, intergalactic fields^{28,29}. These fields could be generated and dispersed by outflows from supermassive black holes^{28,29}. The equivalent magnetic field strength corresponding to the energy density within the cavities is $\sim 100 \mu\text{G}$, considerably larger than the accumulated field strengths in clusters²⁷. Therefore, a plausibly small fraction of the total energy of this one radio outburst alone, if channelled into the cluster's magnetic field, would magnetize the cluster. □

Received 17 August; accepted 12 November 2004; doi:10.1038/nature03202.

1. Fabian, A. C. Cooling flows in clusters of galaxies. *Annu. Rev. Astron. Astrophys.* **32**, 277–318 (1994).
2. Peterson, J. R. et al. High-resolution X-ray spectroscopic constraints on cooling-flow models for clusters of galaxies. *Astrophys. J.* **590**, 207–224 (2003).
3. Fabian, A. C., Mushotzky, R. F., Nulsen, P. E. J. & Peterson, J. R. On the soft X-ray spectrum of cooling flows. *Mon. Not. R. Astron. Soc.* **321**, L20–L24 (2001).
4. McNamara, B. R. et al. Chandra X-Ray Observations of the Hydra A cluster: An interaction between the radio source and the X-ray-emitting gas. *Astrophys. J. Lett.* **534**, 135–138 (2000).
5. Birzan, L. et al. A systematic study of radio-induced X-ray cavities in clusters, groups, and galaxies. *Astrophys. J.* **607**, 800–809 (2004).
6. Fabian, A. C. et al. A deep Chandra observation of the Perseus cluster: shocks and ripples. *Mon. Not. R. Astron. Soc.* **344**, 43–47 (2003).
7. Forman, W. et al. Reflections of AGN outbursts in the gaseous atmosphere of M87. *Astrophys. J.* (in the press).
8. Soker, N. et al. A moderate cluster cooling flow model. *Astrophys. J.* **549**, 832–839 (2001).
9. Nulsen, P. E. J. et al. The cluster-scale AGN outburst in Hydra A. *Astrophys. J.* (submitted).
10. Scheuer, P. A. G. Models of extragalactic radio sources with a continuous energy supply from a central

object. *Mon. Not. R. Astron. Soc.* **166**, 513–528 (1974).

11. Heinz, S., Reynolds, C. S. & Begelman, M. C. X-ray signatures of evolving radio galaxies. *Astrophys. J.* **501**, 126–136 (1998).
12. Condon, J. J. et al. The NRAO VLA sky survey. *Astron. J.* **115**, 1693–1716 (1998).
13. Bicknell, G. V., Dopita, M. A. & O'Dea, C. P. Unification of the radio and optical properties of gigahertz peak spectrum and compact steep-spectrum radio sources. *Astrophys. J.* **485**, 112–124 (1997).
14. De Young, D. S. On the relation between Fanaroff–Riley types I and II radio galaxies. *Astrophys. J. Lett.* **405**, 13–16 (1993).
15. Donahue, M., Stocke, J. T. & Gioia, I. M. Distant cooling flows. *Astrophys. J.* **385**, 49–60 (1992).
16. Churazov, E. et al. Evolution of buoyant bubbles in M87. *Astrophys. J.* **554**, 261–273 (2001).
17. Churazov, E. et al. Cooling flows as a calorimeter of active galactic nucleus mechanical power. *Mon. Not. R. Astron. Soc.* **332**, 729–734 (2002).
18. Edge, A. C. The detection of molecular gas in the central galaxies of cooling flow clusters. *Mon. Not. R. Astron. Soc.* **328**, 762–782 (2001).
19. McNamara, B. R. & O'Connell, R. W. Star formation in cooling flows in clusters of galaxies. *Astron. J.* **98**, 2018–2043 (1989).
20. Donahue, M. & Stocke, J. T. ROSAT observations of distant clusters of galaxies. *Astrophys. J.* **449**, 554–566 (1995).
21. Wu, K. K. S., Fabian, A. C. & Nulsen, P. E. J. Non-gravitational heating in the hierarchical formation of X-ray clusters. *Mon. Not. R. Astron. Soc.* **318**, 889–912 (2000).
22. Markevitch, M. The LX–T relation and temperature function for nearby clusters revisited. *Astrophys. J.* **504**, 27–34 (1998).
23. Gebhardt, K. et al. A relationship between nuclear black hole mass and galaxy velocity dispersion. *Astrophys. J. Lett.* **539**, 13–16 (2000).
24. Stocke, J. T. et al. The Einstein Observatory extended medium-sensitivity survey. II—The optical identifications. *Astrophys. J. Suppl.* **76**, 813–874 (1991).
25. Ferrarese, L. & Merritt, D. A fundamental relation between supermassive black holes and their host galaxies. *Astrophys. J.* **539**, L9–L12 (2000).
26. Heckman, T. M. et al. Present-day growth of black holes and bulges: the SDSS perspective. *Astrophys. J.* **613**, 109–118 (2004).
27. Clarke, T. E., Kronberg, P. P. & Böhringer, H. A new radio–X-ray probe of galaxy cluster magnetic fields. *Astrophys. J. Lett.* **547**, 111–114 (2001).
28. Kronberg, P. P. et al. Magnetic energy of the intergalactic medium from galactic black holes. *Astrophys. J.* **560**, 178–186 (2001).
29. Furlanetto, S. R. & Loeb, A. Intergalactic magnetic fields from quasar outflows. *Astrophys. J.* **556**, 619–634 (2001).

Acknowledgements B.R.McN. thanks G. Evrard, D. De Young, M. Sharma and J. Shields for discussions. The National Radio Astronomy Observatory is operated by Associated Universities Inc., under cooperative agreement with the National Science Foundation. This work was supported by a NASA Long Term Space Astrophysics grant, a Chandra Archival Research grant, a Chandra Guest Observer grant, and a contract from the Department of Energy through the Los Alamos National Laboratory.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.R.McN. (mcnamarb@ohio.edu).

Quantized conductance atomic switch

K. Terabe^{1,2}, T. Hasegawa^{1,2,3}, T. Nakayama^{1,2,3} & M. Aono^{1,2,3}

¹Nanomaterials Laboratories, National Institute for Materials Science, 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan

²ICORP-SORST/Japan Science and Technology Agency, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan

³RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

A large variety of nanometre-scale devices have been investigated in recent years^{1–7} that could overcome the physical and economic limitations of current semiconductor devices⁸. To be of technological interest, the energy consumption and fabrication cost of these 'nanodevices' need to be low. Here we report a new type of nanodevice, a quantized conductance atomic switch (QCAS), which satisfies these requirements. The QCAS works by controlling the formation and annihilation of an atomic bridge at the crossing point between two electrodes. The wires are spaced approximately 1 nm apart, and one of the two is a solid electrolyte

wire from which the atomic bridges are formed. We demonstrate that such a QCAS can switch between 'on' and 'off' states at room temperature and in air at a frequency of 1 MHz and at a small operating voltage (600 mV). Basic logic circuits are also easily fabricated by crossing solid electrolyte wires with metal electrodes.

Mechanical relay switches were the main devices at the dawn of the computer age⁹. However, owing to limitations on their downsizing and operating speed, other devices have replaced them—first vacuum tubes, and then semiconductor devices. We have recently overcome the problems of mechanical switches by using atomic mechanics, which enables us to reduce the size of the device to the nanoscale while keeping the switching speed as fast as present-day electronic devices. Very small energy consumption is also promised, because the device works with a very small bias voltage (for example, 10 mV).

We found that a silver nano-protrusion can be formed at the surface of a silver sulphide (Ag_2S) crystal, which is a mixed ionic and electronic conductor¹⁰, by a solid electrochemical reaction due to a tunnelling current^{11,12}. Growth and shrinkage of the silver protrusion can be controlled by changing the polarity of the bias voltage applied to the material. The QCAS consists of a fixed Ag_2S electrode and a Pt electrode, with a spacing of about 1 nm. A grown silver protrusion forms an atomic bridge between the two electrodes when a positive bias is applied to the Ag_2S electrode. The conductance between the two electrodes then becomes very high, which means the device is switched on. By applying a negative bias to the Ag_2S electrode, the silver protrusion shrinks, thus breaking the atomic bridge, which switches the device off. At the early stage of this study, a scanning tunnelling microscope (STM) was used to make the 1-nm gap between the two electrodes¹³. However, use of a mechanical positioning system (for example, an STM) makes it difficult to use the QCAS as a component of actual devices.

Recently, we found that a QCAS can be formed at each crossing point when a Ag_2S -coated Ag wire is crossed by a Pt wire (Fig. 1a). The 'crossbar' structure is convenient for integrating switches to be used in actual devices¹⁴, and we report here basic circuits using this structure. It would also produce a storage capacity of 2.5 Gbit cm^{-2} using the atomic switch even at the present level of miniaturization of the semiconductor industry. Since the atomic switch itself is

much smaller than this level of miniaturization, much higher density could be made available by improving the miniaturization technology. The crossbar structure has been fabricated with a conventional nanofabricating method, which uses electron-beam lithography and related techniques such as metal deposition and lift-off. That is, Ag_2S -coated Ag wires are formed first by sulphurizing Ag wires. The sulphurization was done at 80°C for 5 min in an ultrahigh vacuum by introducing sulphur vapour using a sulphur valved cracker cell. Then, Pt wires are formed across the Ag_2S -coated Ag wires.

The 1-nm gap, which is the key structure of the atomic switch, is made as follows. A 1-nm-thick Ag layer is deposited on the Ag_2S -coated Ag wires at the places where these wires are to be crossed by the Pt wires—this 1-nm-thick Ag layer is deposited before the formation of the Pt wires. Therefore, the switch is formed in the on state (Fig. 1b). Then, the switch is turned off by applying a certain positive bias voltage to the Pt electrode. Because of the larger amount of Ag atoms to be ionized for incorporating into the Ag_2S crystal in this first turning-off process, the switching time of this process is quite long (a few seconds). However, it becomes shorter after this initial operation.

The QCAS works stably at room temperature in a vacuum and even in air. No difference has been observed in the switching properties between operation in a vacuum and in air. We confirmed that it worked more than 10^5 times, and still continued to work showing constant switching properties—for example, threshold bias voltages and resistances both of the on and off states. We also confirmed that it can be formed by using other solid electrolytes such as Cu_2S , although we will only show here the results for the Ag_2S QCAS. The QCAS can be operated as fast as electronic devices. For instance, switching at 1 MHz by the atomic switch of the crossbar structure is shown in Fig. 2a. At present, switching times less than $1 \mu\text{s}$ cannot be measured, owing to the capacitance of the thick wires forming the QCAS, but we believe that it could work at 1 GHz, on the basis of the following experimental results on switching time.

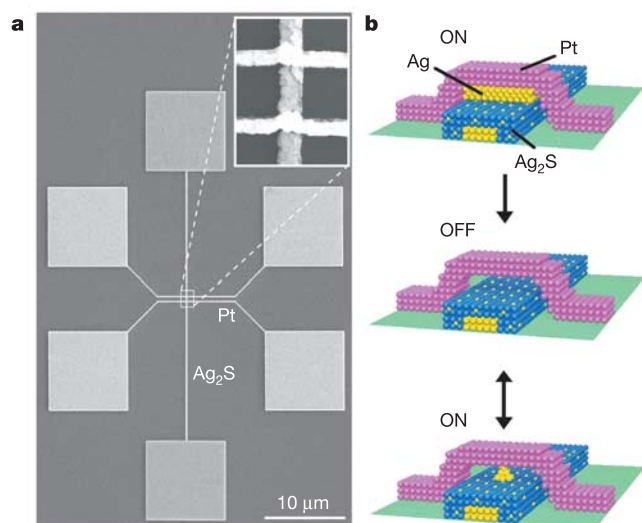


Figure 1 Basics of the QCAS. **a**, SEM image of the QCAS. A QCAS is formed at each crossing point of the 150-nm-wide Ag_2S wire and the two Pt wires of 100 nm width. **b**, Schematic diagrams of the QCAS. As-formed switched-on state (top), switched-off state (middle) and switched-on state after the initial switching-off process (bottom).

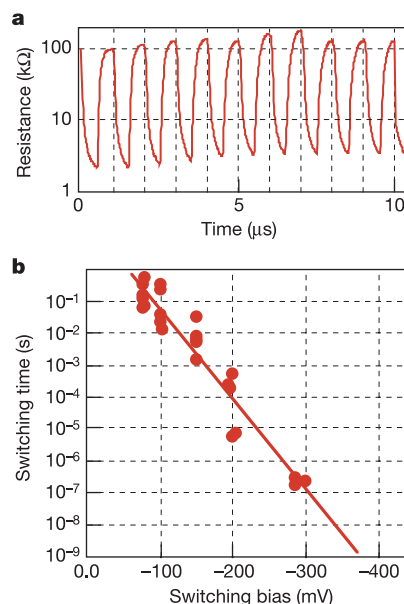


Figure 2 Switching results of the QCAS. **a**, Experimental result of switching at 1 MHz. Alternating switching bias voltages of $\pm 600 \text{ mV}$ were used. **b**, Time taken for the resistance of the QCAS to change from an 'off' resistance ($100 \text{ k}\Omega$) to an 'on' resistance ($12.9 \text{ k}\Omega$, which corresponds to $N = 1 (\times 2e^2/h)$ of the quantized conductance) was measured with respect to the switching bias.

Figure 2b shows the time it takes to switch on the device from the off state with a resistance of 100 k Ω . The result indicates that the switching time decreases exponentially with increasing switching bias voltage in the measured region. This exponential relation suggests that the solid electrochemical reaction, which has some activation energy¹², determines the switching speed. This result suggests that much faster switching could be achieved by using thin wires, which would enable us to apply a bias voltage of higher frequency to the switch. Since silver ions in the Ag₂S crystal are known¹⁵ to be able to hop among the sites when responding to much higher frequencies than those we used, we believe that switching at 1 GHz, for instance, could be achieved with a bias voltage of 0.4 V. Operation as fast as conventional electronic devices could thus be possible, because only a small number of silver atoms have to be moved 1 nm at most for the switching to occur. In other words, downsizing the device to the atomic scale enables its fast operation.

Basic circuits, such as logic gates, have been made using the QCAS. For instance, an AND gate was made using two QCASs and a resistor (Fig. 3a). In the experiment, the crossbar structure having two QCASs (Fig. 1a) was connected to a macro resistor and a measuring system. Two input signals, V_1 and V_2 , are applied to the Pt wires, and the output signal V_{out} was measured at one of the ends of the Ag₂S wire. According to the change in the two input signals, formation and annihilation of atomic bridges were controlled at both QCASs; the changes in the resistances of the two QCASs are shown in Fig. 3a. The operating result is also shown in Fig. 3a, where the output level becomes high only when both inputs are at the high level. By using the resistor, bias voltages effectively applied to the QCASs come into equilibrium soon after switching. In other words, the electrochemical reaction stops soon after each operation. Therefore, too many Ag atoms are not precipitated in each operation, which promises fast operation. The OR gate (Fig. 3b) and NOT gate (Fig. 3c) can be also configured. That all basic logic gates—that is,

AND, OR and NOT—are configurable with QCASs means that any kind of logic circuit can be configured with QCASs.

In the operation of the logic gates, the QCAS is used as an on/off bistable switch. The quantized conductance that the QCAS shows can be also used to make circuits. This conductance is quantized with a unit of $2e^2/h$ (where e is the charge of the electron and h is Planck's constant), as observed in experiments using metal nanowires^{16–24}, as the width of the atomic bridge is comparable to the Fermi length of silver (0.52 nm). By controlling the tip position precisely, switching between quantized conductances has also been demonstrated²⁵. However, the mechanical positioning systems (for example, an STM) employed in these studies make it difficult to use the quantized conductance property of nanowires in actual devices. Recently, a non-mechanical method for forming a nanowire using a liquid electrochemical reaction was proposed^{26,27}, which suggests the possibility of using a nanowire showing quantized conductance in conventional microelectronics.

Switching between quantized conductances of the QCAS is achieved by applying a pulsed bias voltage, which is larger than the threshold bias voltage of the electrochemical reaction^{11,12}. A bias smaller than the threshold bias voltage is used to measure the conductance. Because of the clearness of conductance quantization in a nanowire with a small and completely fixed gap²⁸, the quantized conductance in the QCAS is much more controllable than that in other nanowires^{16–24}. The conductances of multiple QCASs formed in an array are independently controllable. For instance, we controlled the conductances of a 1×2 array of QCASs, as shown in Fig. 4. In the experiment, we changed the conductances at each channel independently from $N = 0$ to 3 by using pulsed bias voltages (here N is the conductance in unit of $2e^2/h$). The total conductance, as calculated from the total current measured at the end of the silver wire, corresponds to the sum of the two conductances, which suggests that the system works as an adder circuit. It also works as a multi-state memory, as it memorizes 16 states simply by using two switches.

Though we have succeeded in demonstrating control of the quantized conductances, further work is needed before these quantized conductances can be used in actual devices. For instance, at this stage, higher numbers of quantized conductances are not as controllable as lower numbers. There is also a problem

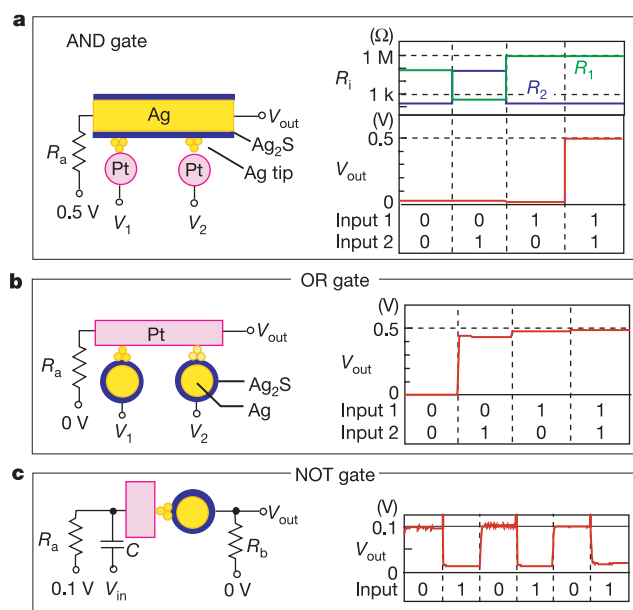


Figure 3 Logic gates configured with QCASs. **a**, Schematic diagram of an AND gate using QCASs (left) and its operating result (right). **b**, Schematic diagram of the OR gate (left) and its operating result (right). **c**, Schematic diagram of the NOT gate (left) and its operating result (right). Resistors R_a (10 k Ω) and R_b (1 k Ω), and a capacitor C (100 pF), are used. V_1 and V_2 are applied as input bias voltages. Input level 1 is 0.5 V for the AND and OR gates, and 1.5 V for the NOT gate. Input level 0 is 0 V for all gates. Input levels were changed every second.

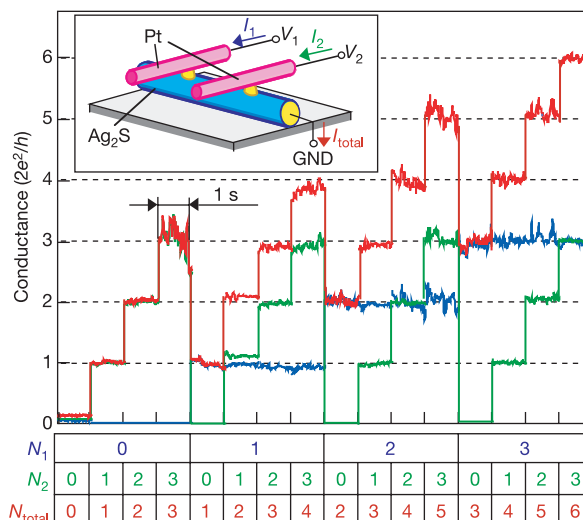


Figure 4 1×2 array of QCASs. The conductances of each channel were changed independently from $N = 0$ ($\times 2e^2/h$) to 3 ($\times 2e^2/h$). 50-ms-long pulsed bias voltages of 200 mV (from 0 to 1), 100 mV (from 1 to 2), 80 mV (from 2 to 3) and -260 mV (from 3 to 0) were used.

with reproducibility even for smaller numbers of quantized conductances, because some switches are found to show less controllability—such as showing a non-integral N of 1.5. This is because the quantized conductance is strongly related to the atomic arrangement of the atomic bridge, which is difficult to control. Therefore, we think that the device structure needs to be improved before practical use of the quantized conductances can be made.

We believe that the simple structure, ease of operation, and the stability and reliability of the QCAS will enable us to use it as an element of future nanodevices, and to make conceptually new electronics that will be part of a new type of computer architecture²⁹. In addition, the QCAS can be used as an element of present-day electronic devices, as the size of the switch itself is already at the atomic scale and it works efficiently at room temperature and in air. □

Received 1 September; accepted 18 November 2004; doi:10.1038/nature03190.

1. Tans, S. J., Verschueren, A. R. M. & Dekker, C. Room-temperature transistor based on a single carbon nanotube. *Nature* **393**, 49–52 (1998).
2. Martel, R., Schmidt, T., Shea, H. R., Hertel, T. & Avouris, Ph. Single- and multi-wall carbon nanotube field-effect transistors. *Appl. Phys. Lett.* **73**, 2447–2449 (1998).
3. Collier, C. P. *et al.* Electronically configurable molecular-based logic gates. *Science* **285**, 391–394 (1999).
4. Joachim, C., Gimzewski, J. K. & Aviram, A. Electronics using hybrid-molecular and mono-molecular devices. *Nature* **408**, 541–548 (2000).
5. Cui, Y. & Lieber, C. M. Functional nanoscale electronic devices assembled using silicon nanowire building blocks. *Science* **291**, 851–853 (2001).
6. Mathur, N. Beyond the silicon roadmap. *Nature* **419**, 573–575 (2002).
7. Duan, X. *et al.* High-performance thin-film transistors using semiconductor nanowires and nanoribbons. *Nature* **425**, 274–278 (2003).
8. Peercy, P. S. The drive to miniaturization. *Nature* **406**, 1023–1026 (2000).
9. Slater, R. *Portraits in Silicon* Ch. 3 & 13 (MIT Press, Cambridge, Massachusetts, 1989).
10. Kudo, T. & Fueki, K. *Solid State Ionics* 137–140 (Kodansha/VCH, Tokyo, 1990).
11. Terabe, K., Nakayama, T., Iyi, N. & Aono, M. in *Proc. 9th Int. Conf. on Precision Engineering* (eds Furukawa, Y., Mori, Y. & Kataoka, T.) 711–716 (The Japan Society for Precision Engineering, Osaka, 1999).
12. Terabe, K., Nakayama, T., Hasegawa, T. & Aono, M. Formation and disappearance of a nanoscale silver cluster realized by solid electrochemical reaction. *J. Appl. Phys.* **91**, 10110–10114 (2002).
13. Terabe, K., Hasegawa, T., Nakayama, T. & Aono, M. Quantum point contact switch realized by solid electrochemical reaction. *Riken Rev.* **37**, 7–8 (2001).
14. Chen, Y. *et al.* Nanoscale molecular-switch devices fabricated by imprint lithography. *Appl. Phys. Lett.* **82**, 1610–1612 (2003).
15. Ohashi, K. & Ohashi, Y. H. Non-linear electrical transport in silver sulfide. *Solid State Ionics* **3/4**, 127–130 (1981).
16. Pascual, J. I. *et al.* Quantum contact in gold nanostructures by scanning tunneling microscopy. *Phys. Rev. Lett.* **71**, 1852–1855 (1993).
17. Olesen, L. *et al.* Quantized conductance in an atom-sized point contact. *Phys. Rev. Lett.* **72**, 2251–2254 (1994).
18. Costa-Kramer, J. L. *et al.* Conductance quantization in nanowires formed between micro- and macroscopic metallic electrodes. *Phys. Rev. B* **55**, 5416–5424 (1997).
19. Ohnishi, H., Kondo, Y. & Takayanagi, K. Quantized conductance through individual rows of suspended gold atoms. *Nature* **395**, 780–785 (1998).
20. Krans, J. M., Van Ruitenbeek, J. M., Fisun, V. V., Yansen, I. K. & de Jongh, L. J. The signature of conductance quantization in metallic point contacts. *Nature* **375**, 767–769 (1995).
21. Hansen, K., Lægsgaard, E., Stensgaard, I. & Besenbacher, F. Quantized conductance in relays. *Phys. Rev. B* **56**, 2208–2220 (1997).
22. Agrait, N., Yeyati, A. L. & Ruitenbeek, J. M. Quantum properties of atomic-sized conductors. *Phys. Rep.* **377** (2–3), 81–279 (2003).
23. Enomoto, A., Kurokawa, S. & Sakai, A. Quantized conductance in Au-Pd and Au-Ag alloy nanocontacts. *Phys. Rev. B* **65**, 125410 (2002).
24. Rodrigues, V., Bettini, J., Rocha, A. R., Rego, L. G. C. & Ugarte, D. Quantum conductance in silver nanowires: correlation between atomic structure and transport properties. *Phys. Rev. B* **65**, 153402 (2002).
25. Smith, D. P. E. Quantum point contact switches. *Science* **269**, 371–373 (1995).
26. Li, C. Z. & Tao, N. J. Quantum transport in metallic nanowires fabricated by electrochemical deposition/dissolution. *Appl. Phys. Lett.* **72**, 894–896 (1998).
27. Xu, B., He, H. & Tao, N. J. Controlling the conductance of atomically thin metal wires with electrochemical potential. *J. Am. Chem. Soc.* **124**, 13568–13575 (2002).
28. Oshima, Y., Mouri, K., Hirayama, H. & Takayanagi, K. Development of a miniature STM holder for study of electronic conductance of metal nanowires in UHV-TEM. *Surf. Sci.* **531**, 209–216 (2003).
29. Heath, J. R., Kuekes, P. J., Snider, G. S. & Williams, R. S. A defect-tolerant computer architecture: opportunities for nanotechnology. *Science* **280**, 1716–1721 (1998).

Acknowledgements We thank M. Kundu and R. Negishi for fabrication of the crossbar-type switches, and T. Tamura for help with measurement of the switching time.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.H. (HASEGAWA.Tsuyoshi@nims.go.jp).

Increasing the conductivity of crystalline polymer electrolytes

Alasdair M. Christie, Scott J. Lilley, Edward Staunton, Yuri G. Andreev & Peter G. Bruce

School of Chemistry, University of St Andrews, St Andrews KY16 9ST, UK

Polymer electrolytes consist of salts dissolved in polymers (for example, polyethylene oxide, PEO), and represent a unique class of solid coordination compounds. They have potential applications in a diverse range of all-solid-state devices, such as rechargeable lithium batteries, flexible electrochromic displays and smart windows^{1–5}. For 30 years, attention was focused on amorphous polymer electrolytes in the belief that crystalline polymer:salt complexes were insulators. This view has been overturned recently by demonstrating ionic conductivity in the crystalline complexes PEO₆:LiXF₆ (X = P, As, Sb); however, the conductivities were relatively low^{6,7}. Here we demonstrate an increase of 1.5 orders of magnitude in the conductivity of these materials by replacing a small proportion of the XF₆[−] anions in the crystal structure with isovalent N(SO₂CF₃)₂[−] ions. We suggest that the larger and more irregularly shaped anions disrupt the potential around the Li⁺ ions, thus enhancing the ionic conductivity in a manner somewhat analogous to the AgBr_{1−x}I_x ionic conductors⁸. The demonstration that doping strategies can enhance the conductivity of crystalline polymer electrolytes represents a significant advance towards the technological exploitation of such materials.

Ionic conductivity in polymer electrolytes was believed to occur in a manner somewhat analogous to gas diffusion through polymer membranes. Segmental motion of the polymer chains continuously creates free volume into which the ions migrate, and this process allows them to progress across the electrolyte^{1–3,9}. Such a view was established by a number of experiments, and denied the possibility of ionic conductivity in crystalline polymers^{10,11}. As a result, attention for the past 30 years has concentrated on amorphous polymer electrolytes, and in particular on the synthesis of new materials with low T_g (glass transition temperature) and hence high levels of segmental motion in order to increase the conductivity¹².

Polymer electrolytes may be prepared as either amorphous or crystalline materials, the latter at certain discrete compositions (ether oxygen to salt ratios). As part of a programme devoted to investigating the structures of the crystalline polymer electrolyte complexes, we solved the crystal structures of the 6:1 complexes PEO₆:LiXF₆ (X = P, As, Sb), and recognized that they had the features necessary to support ionic conductivity in the crystalline state. The structures of the three 6:1 complexes are broadly similar, and are composed of pairs of PEO chains that fold together to form tunnels within which the Li⁺ ions reside: the anions are located outside the tunnels¹³. Subsequently, we demonstrated that these were ionically conducting crystalline polymer electrolytes^{6,7}. Although this was an important development, the conductivities of these materials were low at room temperature (typically 10^{−7} S cm^{−1}). For applications, it is essential to find ways of raising the conductivity. Smart windows and electrochromic displays do not require conductivities as high as those needed for lithium batteries, but values in excess of 10^{−7} S cm^{−1} are essential. We have discovered that it is possible to replace up to 5 mol% of the AsF₆[−] ions in the PEO₆:LiAsF₆ crystal structure by the isovalent N(SO₂CF₃)₂[−] (bis(trifluoromethanesulfonyl)imide or TFSI) ions, resulting in an enhancement in conductivity of 1.5 orders of magnitude. Attempts to replace more AsF₆[−] ions resulted in a

two-phase mixture of the doped crystalline material and liquid $\text{PEO}_6:\text{LiN}(\text{SO}_2\text{CF}_3)_2$.

The 6:1 complexes are stoichiometric materials. The most common approach to raising the conductivity of stoichiometric ceramic conductors is to create vacancies or interstitial ions by doping the materials. However, less obviously the ionic conductivity of stoichiometric ionic conductors may be raised by replacing an ion by a different ion of the same charge. For example, the Ag^+ ion conductivity of crystalline AgBr may be increased by three orders of magnitude by replacing a proportion of the Br^- ions by I^- ions⁸. The substituting ion changes the potential energy of the conducting ion and hence the energetics of defect creation as well as ion mobility. In the present case we have investigated isovalent ion doping in crystalline polymer electrolytes by replacing the AsF_6^- ion in the crystalline complex $\text{PEO}_6:\text{LiAsF}_6$ with the TFSI^- ion. We chose the TFSI^- ion because it is significantly different—in size, shape and charge distribution—from the AsF_6^- ion, and hence should induce differences in the potential energy of the lithium ions while, although larger than AsF_6^- , remaining within dimensions that may be incorporated into the $\text{PEO}_6:\text{LiAsF}_6$ crystal structure, as discussed later.

A series of compositions in which the ether oxygen to cation ratio was maintained at 6:1 but the LiAsF_6 salt was continuously replaced with LiTFSI (that is, compositions corresponding to $\text{PEO}_6:(\text{LiAsF}_6)_{1-x}(\text{LiTFSI})_x$) were prepared for $0 \leq x \leq 1$, as described in the Methods section. As in our previous studies of the pure $\text{PEO}_6:\text{LiAsF}_6$ complexes, methoxy-endcapped PEO was used to ensure chemical homogeneity between the ends and the

bulk of the chains. Furthermore, and again as before, an average molar mass of 1,000 was selected, which yields the highest conductivities for the pure 6:1 phases.

Conductivity isotherms are presented in Fig. 1a, and indicate how the conductivity of the system varies with lithium TFSI content. The variation of conductivity with temperature is shown in Fig. 1b. Two distinct behaviours are evident in Fig. 1a, and these are divided into regions A and B in the figure. Starting with the pure $\text{PEO}_6:\text{LiAsF}_6$, the conductivity rises rapidly by 1.5 orders of magnitude with x up to a maximum composition of around $\text{PEO}_6:(\text{LiAsF}_6)_{0.95}(\text{LiTFSI})_{0.05}$. Thereafter, in region B, further increases in TFSI content result in a much more gradual rise in the conductivity, until pure $\text{PEO}_6:\text{LiTFSI}$ is reached at $x = 1$. Note that $\text{PEO}_6:\text{LiTFSI}$ with PEO of average molar mass 1,000 is a viscous liquid, and is not suitable as a solid separator in lithium batteries despite its high conductivity. This is, of course, different from the high-molecular-weight amorphous $\text{PEO}:\text{LiTFSI}$ materials¹⁴. To understand the processes involved in enhancing the conductivity compared with the pure $\text{PEO}_6:\text{LiAsF}_6$, and especially to understand the very

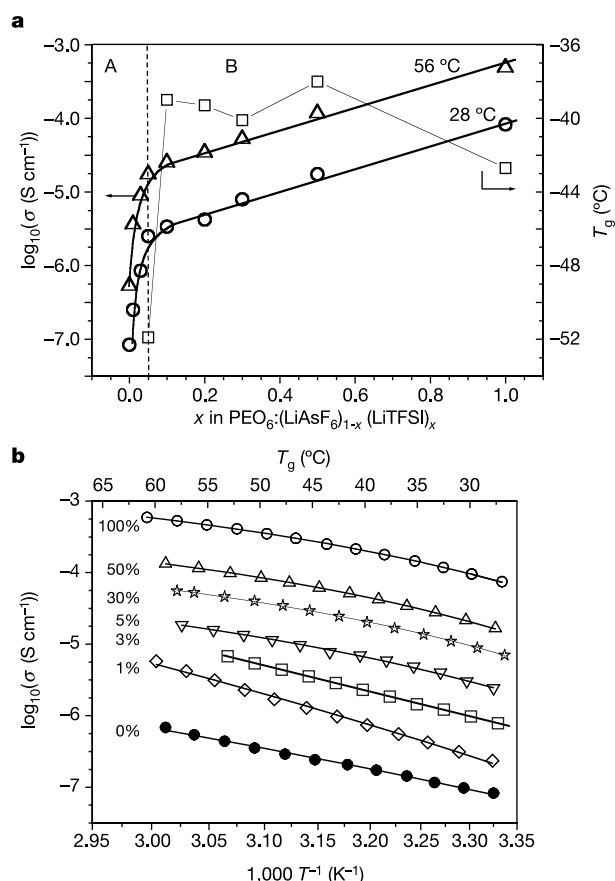


Figure 1 Conductivity of $\text{PEO}_6:(\text{LiAsF}_6)_{1-x}(\text{LiTFSI})_x$. **a**, Conductivity isotherms as a function of x . Regions A and B are defined in the text. **b**, Ionic conductivity as a function of temperature; numbers labelling curves are values of x in mol%. Solid lines represent best least-squares fits to the expressions described in the text.

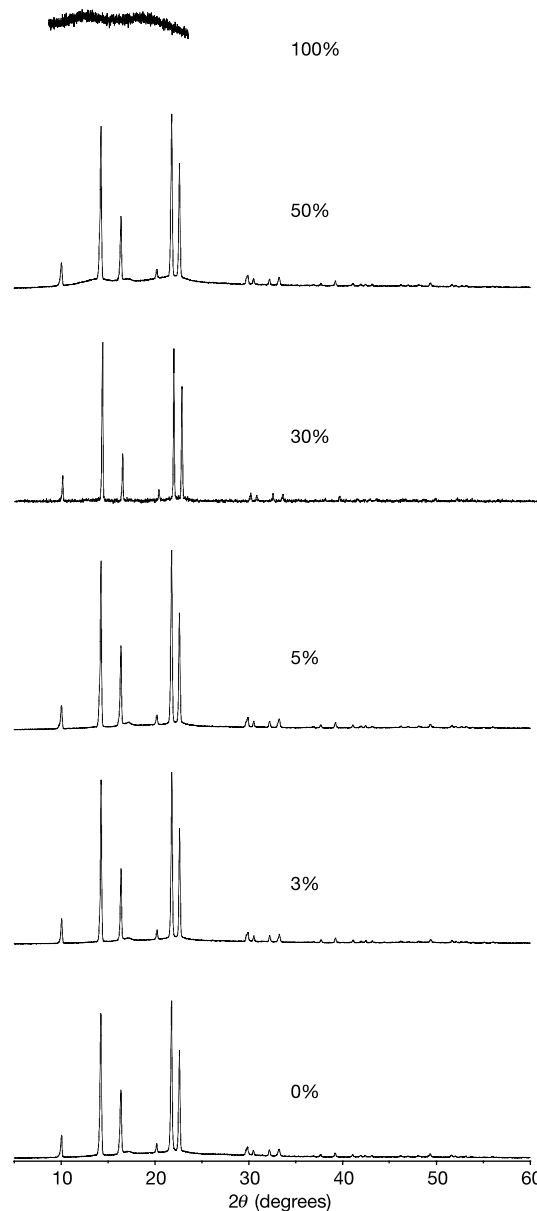


Figure 2 X-ray powder diffraction patterns of $\text{PEO}_6:(\text{LiAsF}_6)_{1-x}(\text{LiTFSI})_x$. Numbers labelling patterns are values of x in mol%.

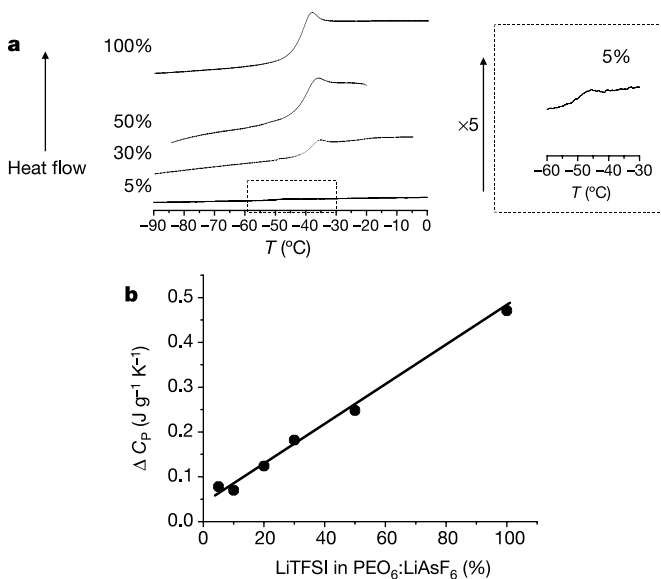


Figure 3 DSC of PEO₆:(LiAsF₆)_{1-x}(LiTFSI)_x. **a**, Sections of the traces showing T_g ; numbers labelling traces are values of x in mol%. Dashed box in left panel shown magnified on right. **b**, Heat capacity change at T_g as a function of x .

different behaviour exhibited in regions A and B of Fig. 1a, it is important to characterize the nature of the phases present in the PEO₆:(LiAsF₆)_{1-x}(LiTFSI)_x system as a function of x .

Powder X-ray diffraction patterns for several compositions are shown in Fig. 2. The amorphous nature of PEO₆:LiTFSI is evident (100%). There is no indication of unreacted LiTFSI salt at any composition (the strongest diffraction peak in the pattern of the salt is located between 19° and 20° in 2θ). Furthermore, there is no significant shift in the peak positions of the PEO₆:LiAsF₆ phase or change in the relative peak intensities with x , indicating that any incorporation of TFSI⁻ into the PEO₆:LiAsF₆ crystal structure must be limited to values below those for which changes in the powder X-ray diffraction pattern would be anticipated. (Simulation of the powder X-ray pattern in which 5% of AsF₆⁻ is replaced by TFSI⁻ shows that the changes in the relative peak intensities are below the detection limit of the powder diffraction method.) As a result, we may conclude that over most of the composition range (region B in

Fig. 1a) a two-phase mixture exists between amorphous PEO₆:LiTFSI and a crystalline phase with the PEO₆:LiAsF₆ structure.

Differential scanning calorimetry (DSC) data were collected for a range of compositions (Fig. 3a). Compositions greater than $x = 0.05$ exhibit a glass transition. The magnitudes of the heat capacity differences associated with this transition (that is, the size of the step associated with T_g in the DSC traces) were extracted using Netzsch software and are presented in Fig. 3b. It is evident that they increase linearly with LiTFSI content, consistent with the view that a two-phase mixture exists between liquid PEO₆:LiTFSI and a crystalline phase based on PEO₆:LiAsF₆, in region B of Fig. 1a. Below $x = 0.05$ there is no evidence of a T_g ; furthermore, at $x = 0.05$, the T_g now shows a sharp drop compared with the T_g values at higher x . The variation in T_g with x observed in Fig. 1a is exactly what is to be expected for an amorphous phase that becomes significantly depleted in salt at $x = 0.05$ (hence T_g drops) compared with compositions at higher values of x . These facts are consistent with the view that TFSI is incorporated into the crystal structure of the PEO₆:LiAsF₆ phase up to a composition limit $x \approx 0.05$; thereafter, a two-phase mixture between crystalline PEO₆:(LiAsF₆)_{0.95}(LiTFSI)_{0.05} and liquid PEO₆:LiTFSI exists. On the basis of this analysis of the phase behaviour, we may now interpret the trends in the conductivity with composition. Region B is straightforward. The crystalline phase is embedded in a highly conducting liquid of composition PEO₆:LiTFSI, and conductivity increases with increasing LiTFSI content because the proportion of the liquid phases increases.

Of much greater interest is region A. Starting from pure PEO₆:LiAsF₆ and increasing the LiTFSI content; replacement of AsF₆⁻ by TFSI⁻ within the 6:1 crystal structure results in a marked increase in conductivity. Ion transport in the crystalline phase is believed to occur owing to the presence of Li⁺ defects within the polymer tunnels that promote Li⁺ ion hopping between sites along the tunnels. The disruption caused by substituting AsF₆⁻ ions by the larger and differently shaped TFSI⁻ ions is expected to disrupt the potential experienced by Li⁺ ions in the region of the TFSI⁻, resulting in an increase in conductivity in a manner analogous to the AgBr_{1-x}I_x ionic conductors described above⁸. Although the size of the TFSI⁻ ion is greater than that of AsF₆⁻, and we do not know its exact conformation in the crystal structure, an analysis of the dimensions available to TFSI⁻ on substituting for AsF₆⁻ indicates that space is available for its incorporation (Fig. 4).

Confirmation that conduction in region A occurs by a hopping mechanism and is different from region B is obtained by least-

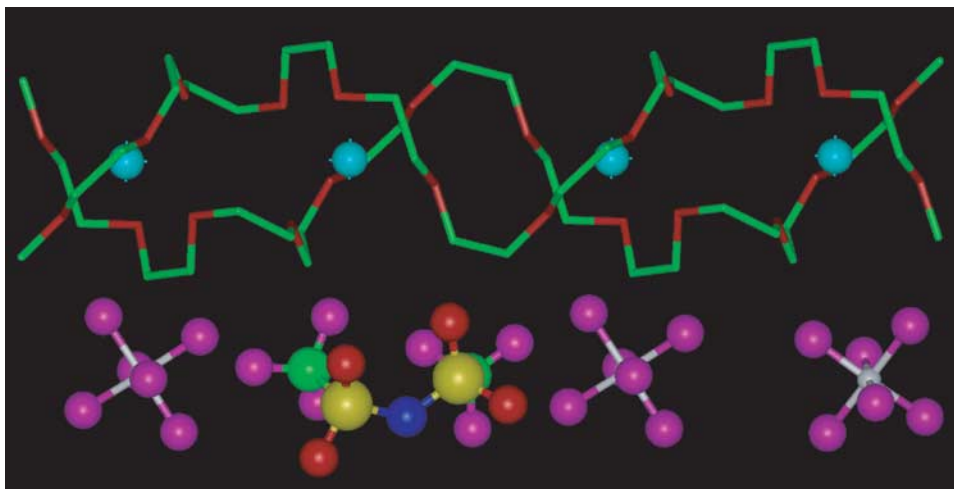


Figure 4 Fragment of the crystal structure of PEO₆:(LiAsF₆)_{1-x}(LiTFSI)_x. Figure shows the substitution of the AsF₆⁻ ion by TFSI⁻. Light blue, lithium; white, arsenic; purple, fluorine; dark blue, nitrogen; yellow, sulphur; green, carbon; red, oxygen.

squares fitting to the temperature-dependent conductivities in Fig. 1b. The 0%, 1% and 3% materials exhibit linear $\log \sigma$ versus $1/T$ plots, and are well described by an Arrhenius expression. Fitting such an expression to the data for these compositions gives activation energies of 55, 83 and 70 kJ mol⁻¹ respectively. An Arrhenius expression may also be used to describe the 5% doped material, although the use of a Vogel–Tamman–Fulcher (VTF) expression $\sigma = \sigma_0 T^{-1/2} \exp(-B/(T - T_0))$ gives a slightly better fit. When we take into account the fact that the VTF expression involves a 50% increase in the number of variables used to fit the conductivity data, compared with the Arrhenius expression, the improvement of the fit is marginal. The onset of such a marginal curvature at 5% is exactly what is expected for a composition on the border between region A and region B, and hence is further confirmation of the doping mechanism. With increasing LiTFSI content in region B the $\log \sigma$ versus $1/T$ plots become more curved, and can only be described by a VTF equation. These trends reinforce the interpretation that region A is dominated by conduction in a doped crystalline material. In parallel with the discovery of ionic conductivity in the crystalline polymer:salt complexes, developments are taking place in the ionic conductivity of plastic crystalline materials^{15,16}. Although different, the role of defects in conduction is important in both these classes of ionic conductors.

We have demonstrated that it is possible to raise the conductivity of crystalline polymer electrolyte by isovalent doping, and we anticipate that variations of this strategy may lead to other dopants and yet higher conductivity. □

Methods

Samples of PEO₆(LiAsF₆)_{1-x}(LiTFSI)_x were prepared by dissolving appropriate quantities of LiAsF₆ (ABCR, 99.8%), LiTFSI (3M) and -OCH₃ terminated poly(ethylene oxide) with an average molar mass of 1,000 (Fluka, ≥99.5%) together in dry acetonitrile (Aldrich, 99.8%). All constituents were dried before use, and all manipulations were carried out in a high-integrity argon-filled MBraun glove box. After dissolution, the solutions were transferred into glass vials and the solvent allowed to evaporate slowly. The resulting complexes were dried at room temperature under dynamic vacuum for at least 24 h. DSC was carried out using a Netzsch DSC 204 Phoenix with heating and cooling rates of 5° min⁻¹.

Powder X-ray diffraction data were collected at room temperature, in transmission mode, using a STOE STADI/P diffractometer with Cu K_{α1} radiation and a position-sensitive detector. The polymer electrolyte samples were sealed inside glass capillaries, and data were collected with a step size of 0.02° in 2θ.

For conductivity measurements, a sample of each complex was pressed at room temperature between two 0.025-mm-thick stainless-steel disks. These self-supporting disks were placed into two-electrode cells that were, in turn, sealed inside argon-filled cans for removal from the glove box. Each can was placed into an oil bath equipped with a Haake EK30 cooler and a Haake DL30 temperature controller connected to a PC. All internal cell temperatures were monitored using K-type thermocouples also connected to the PC. Conductivity data were obtained using a.c. impedance measurements carried out with a Solartron 1255 frequency response analyser coupled with a Solartron 1286 electrochemical interface. A perturbation voltage of 10 mV was applied over the frequency range 500 kHz–1 Hz. All instruments were connected to the PC, so that the temperature sweep and each a.c. impedance measurement were driven by custom software. Before measurements were made at each temperature, a 2-h equilibration period was enforced after the internal cell temperature of each cell had reached a steady state.

Received 27 September; accepted 11 November 2004; doi:10.1038/nature03186.

- Scrosati, B. (ed.) *Applications of Electroactive Polymers* (Chapman & Hall, London, 1993).
- Bruce, P. G. *Solid State Electrochemistry* (Cambridge Univ. Press, Cambridge, 1995).
- Gray, F. M. *Polymer Electrolytes* (RSC Materials Monographs, The Royal Society of Chemistry, Cambridge, 1997).
- Tarascon, J.-M. & Armand, M. Issues and challenges facing rechargeable lithium batteries. *Nature* **414**, 359–367 (2001).
- Barnes, A. *et al.* Towards a 'smart window' for microwave applications. *Smart Mater. Struct.* **7**, 752–758 (1998).
- Gadjourova, Z., Andreev, Y. G., Tunstall, D. P. & Bruce, P. G. Ionic conductivity in crystalline polymer electrolytes. *Nature* **412**, 520–523 (2001).
- Stoeva, Z., Martin-Litas, I., Staunton, E., Andreev, Y. G. & Bruce, P. G. Ionic conductivity in the crystalline polymer electrolytes PEO₆LiX₆, X = P, As, Sb. *J. Am. Chem. Soc.* **125**, 4619–4626 (2003).
- Shahi, K. & Wagner, J. B. Jr. Fast ion transport in silver halide solid solutions and multiphase mixtures. *Appl. Phys. Lett.* **37**, 757–759 (1980).
- Ratner, M. A. in *Polymer Electrolytes Reviews – 1* (eds MacCallum, J. R. & Vincent, C. A.) 173–236 (Elsevier Applied Science, London, 1987).
- Berthier, C. *et al.* Microscopic investigation of ionic conductivity in alkali metal salts – poly(ethylene oxide) adducts. *Solid State Ionics* **11**, 91–95 (1983).

- Ratner, M. A. & Shriver, D. F. Ion-transport in solvent-free polymers. *Chem. Rev.* **88**, 109–124 (1988).
- Angell, C. A., Liu, C. & Sanchez, E. Rubbery solid electrolytes with dominant cationic transport and high ambient conductivity. *Nature* **362**, 137–139 (1993).
- Gadjourova, Z., Martin y Marero, D., Andersen, K. H., Andreev, Y. G. & Bruce, P. G. Structures of the polymer electrolyte complexes PEO₆LiX₆ (X = P, Sb), determined from neutron powder diffraction data. *Chem. Mater.* **13**, 1282–1285 (2001).
- Sylla, S., Sanchez, J.-Y. & Armand, M. Electrochemical study of linear and crosslinked POE-based polymer electrolytes. *Electrochim. Acta* **37**, 1699–1701 (1992).
- MacFarlane, D. R., Huang, J. H. & Forsyth, M. Lithium-doped plastic crystal electrolytes exhibiting fast ion conduction for secondary batteries. *Nature* **402**, 792–794 (1999).
- Forsyth, M., Huang, J. & MacFarlane, D. R. Lithium doped N-methyl-N-ethylpyrrolidinium bis(trifluoromethanesulfonyl) amide fast-ion conducting plastic crystals. *J. Mater. Chem.* **10**, 2259–2265 (2000).

Acknowledgements P.G.B. is indebted to The Royal Society, the EU and the EPSRC for financial support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.G.B. (p.g.bruce@st-and.ac.uk).

Rapid stepwise onset of Antarctic glaciation and deeper calcite compensation in the Pacific Ocean

Helen K. Coxall^{1,†}, Paul A. Wilson¹, Heiko Pälike^{2,†}, Caroline H. Lear^{3,†} & Jan Backman²

¹Southampton Oceanography Centre, School of Ocean and Earth Science, European Way, Southampton SO14 3ZH, UK

²Geology and Geochemistry, University of Stockholm, S-10691 Stockholm, Sweden

³Institute of Marine and Coastal Sciences, Rutgers University, 71 Dudley Road, New Brunswick, New Jersey 08901 USA

* Present addresses: Graduate School of Oceanography, University of Rhode Island, Bay Campus, South Ferry Rd, Narragansett, Rhode Island 02882, USA (H.K.C.); Southampton Oceanography Centre, School of Ocean and Earth Science, European Way, Southampton SO14 3ZH, UK (H.P.); School of Earth, Ocean and Planetary Sciences, PO Box 914, Cardiff University, Cardiff CF10 3YE, UK (C.H.L.).

The ocean depth at which the rate of calcium carbonate input from surface waters equals the rate of dissolution is termed the calcite compensation depth. At present, this depth is ~4,500 m, with some variation between and within ocean basins. The calcite compensation depth is linked to ocean acidity, which is in turn linked to atmospheric carbon dioxide concentrations and hence global climate¹. Geological records of changes in the calcite compensation depth show a prominent deepening of more than 1 km near the Eocene/Oligocene boundary (~34 million years ago)² when significant permanent ice sheets first appeared on Antarctica^{3–6}, but the relationship between these two events is poorly understood. Here we present ocean sediment records of calcium carbonate content as well as carbon and oxygen isotopic compositions from the tropical Pacific Ocean that cover the Eocene/Oligocene boundary. We find that the deepening of the calcite compensation depth was more rapid than previously documented and occurred in two jumps of about 40,000 years each, synchronous with the stepwise onset of Antarctic ice-sheet growth. The glaciation was initiated, after climatic preconditioning⁷, by an interval when the Earth's orbit of the Sun favoured cool summers. The changes in oxygen-isotope composition across the Eocene/Oligocene boundary are too large to be explained by Antarctic ice-sheet growth alone and must therefore also indicate contemporaneous global cooling and/or Northern Hemisphere glaciation.

The pattern of post-Eocene climate change revealed by benthic $\delta^{18}\text{O}$ records is one of abrupt (<1 million year, Myr) increases, superimposed on longer-term $\delta^{18}\text{O}$ increases reflecting a combination of global cooling and ice growth^{3–6}. Most prominent of the abrupt increases is 'Oi-1' near the Eocene/Oligocene boundary^{3–6}. Oi-1 is thought to mark the initiation of major permanent Cenozoic ice-sheets on Antarctica but its cause is widely debated. One hypothesis is that Oi-1 was triggered by the opening of Southern Ocean gateways³; another is that it was caused by a threshold response to long-term Cenozoic decline in atmospheric carbon dioxide levels⁷. The Eocene/Oligocene transition also shows a marked calcite compensation depth (CCD) increase in classic Deep Sea Drilling Project (DSDP) records (Fig. 1 inset). However, the timing and duration of CCD increase, the amplitude and implied ice budget of Oi-1 and the relationship between these two events are poorly constrained. Most DSDP and Ocean Drilling Program (ODP) sites spanning the Eocene/Oligocene boundary are afflicted by condensation horizons and hiatuses attributed to increases in ocean circulation vigour and glacioeustatic sea-level fall associated with Antarctic ice-sheet growth^{3–5}. Thus, previous interpretations of this key interval are heavily based on records from a handful of mid- to high-latitude sites, none of which are complete, or from the Pacific—the world's largest ocean.

ODP Leg 199 recovered multiple Eocene/Oligocene boundary sections in the tropical Pacific Ocean with unprecedented magneto- and cyclostratigraphic age control⁸. The Eocene/Oligocene tran-

sition is instantly recognizable in these strata by up-section shifts from opal-rich to carbonate-rich sediments. We have developed a new chronology for these sections, based on detailed correlation of geological data to astronomically calculated variations of Earth's orbit and solar insolation (ref. 9; the new astronomical solution is available from <http://www.imcce.fr/Equipes/ASD/insola/earth/earth.html>). ODP Site 1218 provides the best record across the Eocene/Oligocene boundary. To reconstruct CCD changes we determined bulk weight per cent CaCO_3 and CaCO_3 mass accumulation rate (MAR) by direct measurement on discrete samples and by regression from whole-core analyses of physical properties (Fig. 1). CaCO_3 MAR dominates bulk sediment MAR and our records show strong variations on orbital timescales. The sharp increase in CaCO_3 around 34 Myr ago (Fig. 1), and the correlative appearance of CaCO_3 in deeper water sites⁸, suggest that the classic DSDP record (Fig. 1 inset) accurately captures the magnitude of the Eocene/Oligocene CCD increase (≥ 1 km) in the Pacific Ocean. But the classic DSDP record does not constrain the timing and duration of the shift to better than a 2–3-Myr interval², and subsequent studies suggest that it occurred gradually (over several million years)^{5,10}. Our new data show that CCD increase took place an order of magnitude faster than this in the Pacific and not as a single event but in two steps (~ 40 kyr each) separated by an intermediate plateau (~ 200 kyr, Fig. 1).

To investigate the cause of Eocene/Oligocene CCD increase and its relation to changes in global climate and carbon cycling, we

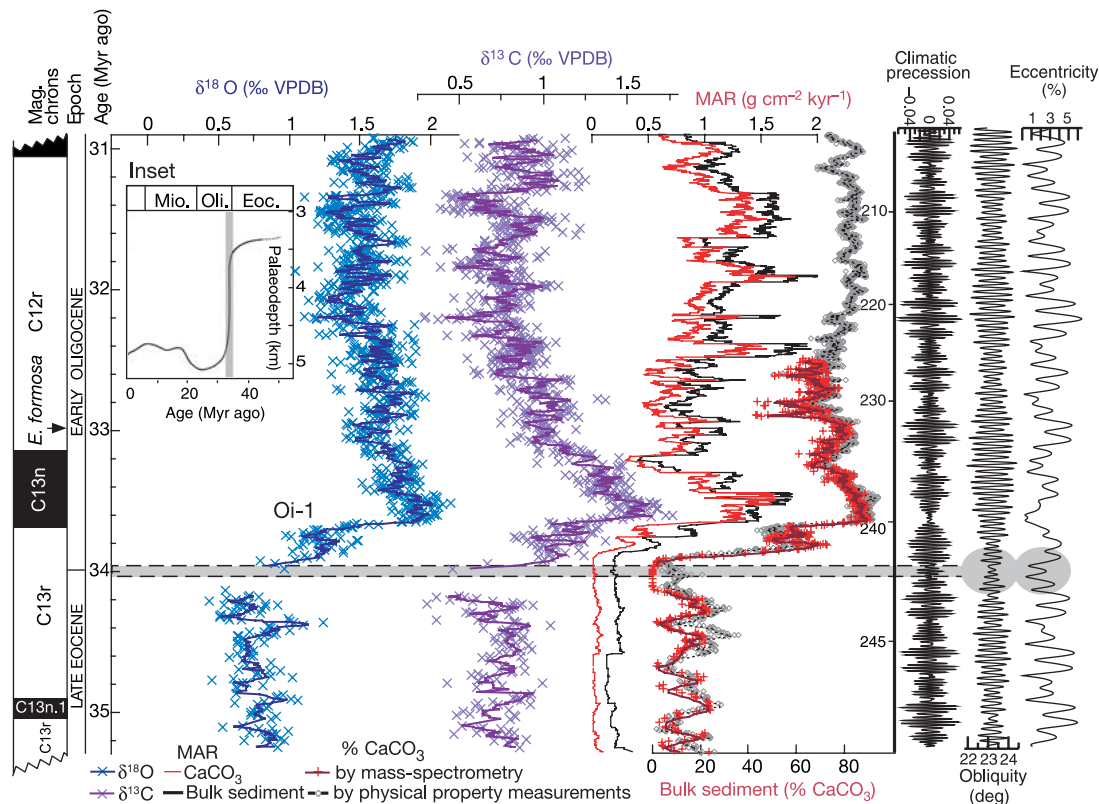


Figure 1 Palaeoceanographic records showing changes in global climate and ocean chemistry for the Eocene/Oligocene transition. The inset shows published² CCD for the equatorial Pacific Ocean 50 Myr ago to the present, from classic Deep Sea Drilling Project sites. This published record shows a 1-km deepening near the Eocene/Oligocene boundary but the timing and duration of this shift is poorly constrained (shading = uncertainty of ~ 3 Myr). The main figure shows new high-resolution data (~ 35.5 to 31.5 Myr ago) from ODP site 1218 ($8^\circ 53.38' \text{ N}$; $135^\circ 22.00' \text{ W}$, water

depth = 4,862 m; 34 Myr ago palaeolatitude ~ 0 to 2° N ; palaeodepth $\approx 3,800$ m), showing that CCD increase (increase in CaCO_3) occurred (1) faster than previously documented; (2) in two 40-kyr steps, (3) synchronously with the stepwise onset of major permanent Cenozoic Antarctic ice-sheets ($\delta^{18}\text{O}$ increase in benthic foraminiferal calcite) and (4) during an eccentricity minimum and low-amplitude obliquity change (grey shading) favouring cool summers. Benthic $\delta^{13}\text{C}$ also shows a stepwise increase. $\delta^{18}\text{O}$ = blue; $\delta^{13}\text{C}$ = purple; CaCO_3 = red.

analysed $\delta^{18}\text{O}$ and $\delta^{13}\text{C}$ in benthic foraminifera (Fig. 1). Our records show that increases in $\delta^{18}\text{O}$ and $\delta^{13}\text{C}$ documented in mid- and high-latitude South Atlantic⁵ and Southern Ocean¹¹ sites also occur in the Pacific, confirming that these are truly global signals with stratigraphic utility. The contemporaneous occurrence of ice-rafted debris, and a shift from clay mineral assemblages dominated by smectite to those dominated by illite and chlorite in the Southern Ocean, suggest that the first major permanent Cenozoic ice sheets appeared on Antarctica in the early Oligocene^{11,12}. Our new records are of the highest resolution (up to 2 kyr) yet achieved for this interval and shed new light on this important event.

Our isotope records show pronounced increases in $\delta^{18}\text{O}$ and $\delta^{13}\text{C}$ that are synchronous with CCD increase, demonstrating that the transition from a relatively deglaciated climate state in the latest Eocene to a climate state with well-developed ice sheets on Antarctica in earliest Oligocene time was completed within 300 kyr (Fig. 1). Remarkably, the pattern of isotopic increase has the same distinctive stepwise form as our %CaCO₃ series, with most of the $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ shift taking place in two 40-kyr-long 'steps'. The two-step Eocene/Oligocene transition in $\delta^{18}\text{O}$ occurs in lock-step with that in CaCO₃ MAR, but the two-step increase in $\delta^{13}\text{C}$ appears to occur slightly later (a <10 kyr lag). Furthermore, $\delta^{13}\text{C}$, $\delta^{18}\text{O}$ and CaCO₃ MAR all show a distinctive 'overshoot' of typical early Oligocene values during their earliest Oligocene maxima (Fig. 1). The initiation of step-change in our records occurs during an interval of low eccentricity and low-amplitude change in obliquity, conditions favouring dampened seasonality (Fig. 1). This observation is consistent with the view⁷ that it was the prolonged absence of warm summers, inhibiting summer snow melt, not the occurrence of cool winters favouring accumulation, that was

important for establishing the first major Cenozoic ice sheets on Antarctica.

The apparent contradiction between our findings and the increased seasonality seen across the Eocene/Oligocene boundary in fish otolith records¹³ probably reflects the lack of otolith data from the transition interval of low seasonality because our records demonstrate that, once the Antarctic ice sheet was established, Early Oligocene global climate was highly sensitive to increasing power (seasonality) in the astronomical series (Fig. 1). In other words, the otolith study documents the effect of Antarctic glaciation, whereas our data document its cause. Cross-spectral analyses of our two Oligocene isotope series (~33.25 to 31 Myr ago) show a phase lag of $\delta^{18}\text{O}$ with respect to $\delta^{13}\text{C}$ corresponding to about 8 kyr in the 40-kyr band (Fig. 2), suggesting that the response of the global carbon cycle to Earth's obliquity helped to force changes in early Oligocene climate.

Our data indicate that Earth's orbital configuration was the ultimate trigger for Oi-1 and the pacemaker for ice-sheet growth. Yet some other conditioning factor must have been important, because there is no evidence to suggest that the low-eccentricity obliquity 'node' conditions at 34 Myr ago are more extreme than those occurring every 2.4 Myr (the eccentricity minimum) and 1.2 Myr (the obliquity minimum) during the past 40 Myr, and

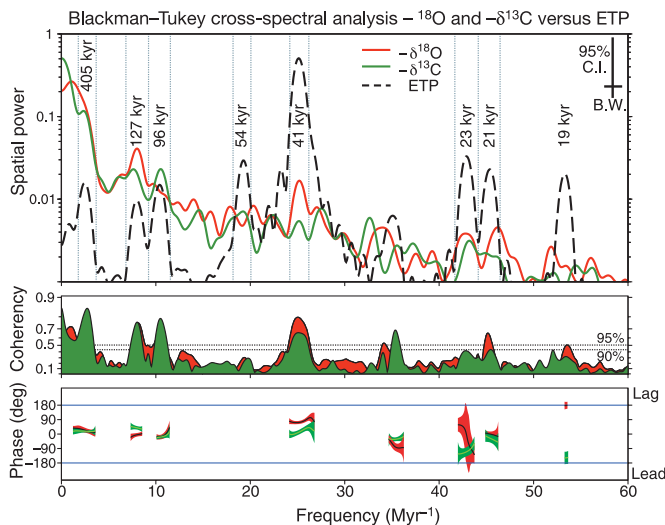


Figure 2 Spectral analysis of ODP site 1218 benthic stable isotopes $\delta^{18}\text{O}$ = red; $\delta^{13}\text{C}$ = green) and astronomical solution⁹ (black dashed line). All astronomical frequencies are encoded within the two Oligocene isotope series, with power concentrated at obliquity ($\delta^{18}\text{O}$) and 400-kyr eccentricity ($\delta^{13}\text{C}$) frequencies. Cross-spectral analyses indicate that both isotope series are coherent with eccentricity and obliquity above the 95% confidence level, and with elements of climatic precession between the 90% to 95% confidence level. Error bars indicate 95% confidence levels of phase estimates. ETP = mix of eccentricity, tilt (obliquity) and climatic precession, adjusted to simulate the strong eccentricity signal in our data. The phase relationship between the strong obliquity component of the $\delta^{18}\text{O}$ data and the weaker component in the $\delta^{13}\text{C}$ data suggests a lag of ~8 kyr of $\delta^{18}\text{O}$ in the ~40-kyr band (at the ~25 Myr⁻¹ frequency). For the high-amplitude 405 kyr and ~100 kyr eccentricity peaks there is no phase difference between the isotope series. B.W., band width.

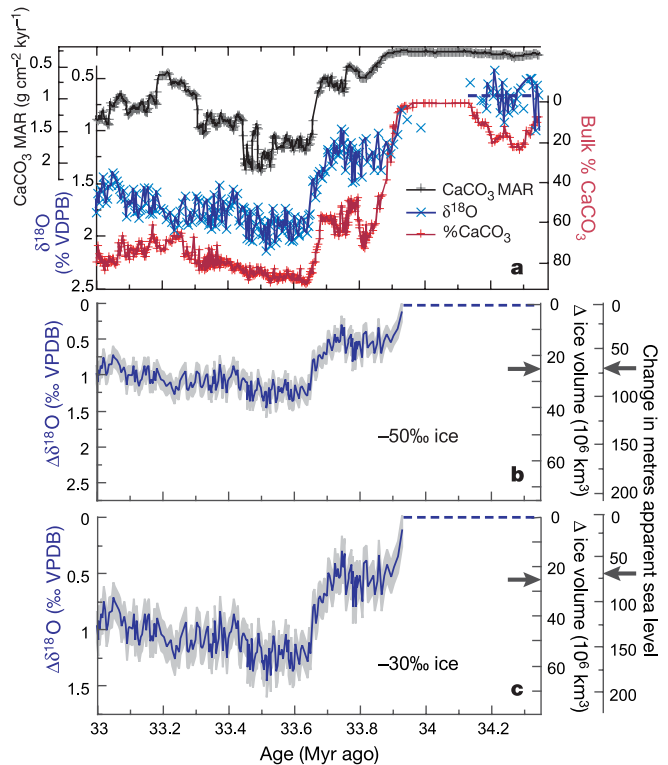


Figure 3 ODP site 1218 records expanded from Fig. 1 and implications of the large $\delta^{18}\text{O}$ increase. **a**, Eocene/Oligocene records of $\delta^{18}\text{O}$ and CCD (%CaCO₃ and CaCO₃ MAR) from Fig. 1. Dashed blue line shows 'base line' mean benthic $\delta^{18}\text{O}$ for latest Eocene time (last 400 kyr). **b, c**, The early Oligocene increase in $\delta^{18}\text{O}$ relative to this baseline ($\Delta\delta^{18}\text{O}$) and associated estimated change in global ice budget and metres apparent sea level¹⁷ (m ASL) assuming all of $\Delta\delta^{18}\text{O}$ is attributable to increased ice volume (as suggested by Mg/Ca records, see text). Blue, best estimate; grey, estimated uncertainty for sea level change associated with one standard deviation about the Eocene base line. Panel **b** assumes, conservatively, an isotopic composition of Oligocene ice equal to the average for the Antarctic today (~50‰). Panel **c** assumes Oligocene ice is ~30‰ (see text). Arrows indicate modern Antarctic ice volume (~25.4 × 10⁶ km³) and ASL fall (70 m) estimated¹⁷ for the Eocene/Oligocene transition by sequence stratigraphy.

possibly back to 50 Myr ago^{9,14}. Results of a recent global climate model (GCM) experiment⁷ suggest that this conditioning factor was a long-term decline in Cenozoic atmospheric CO₂ levels (Eocene/Oligocene ice-sheet threshold ~2.8 to 3 times the pre-anthropogenic value). In fact, the sudden jumps in ice volume implied by the steps in our $\delta^{18}\text{O}$ data are strikingly similar in form to, but three times larger than, those simulated in ref. 7. The steps in the $\delta^{18}\text{O}$ data substantiate model predictions⁷ of a rapid, non-steady mode of ice-sheet growth in response to mass-balance feedback effects associated with ice-sheet height and coalescence.

Two main factors explain the large amplitude of Oi-1 in our data ($1.5 \pm 0.1\text{‰}$) relative to that modelled ($\sim 0.5\text{‰}$)⁷. First, the model considers only Antarctic ice sheets and these are constrained by the coastline, whereas our data reflect continental ice sheets globally and these may expand by seaward advance over newly exposed continental shelf in response to glacioeustatic sea-level fall (grounding line advance). Second, the model considers only ice volume-driven $\delta^{18}\text{O}$ increase, whereas our data may incorporate an additional temperature component (cooling). Eocene/Oligocene records of Mg/Ca in deep-sea foraminifera indicate no cooling, which suggests^{6,15,16}, controversially¹³, that all of the $\delta^{18}\text{O}$ increase is attributable to ice growth. In Fig. 3 we use our $\delta^{18}\text{O}$ data from the central Pacific as a globally representative record to test this suggestion by calculating implied global ice volumes.

Assuming, conservatively, that no pH/[CO₃²⁻] effect associated with CCD increase acts to suppress $\delta^{18}\text{O}$ increase across the Eocene/Oligocene transition and that the average isotopic composition of Antarctic ice was as extreme as today (-50‰) we calculate a global Early Oligocene ice-volume maximum and apparent sea-level¹⁷ (ASL) fall that are about 1.6 times both the modern total Antarctic ice budget ($25.4 \times 10^6 \text{ km}^3$) (ref. 18) and the sequence stratigraphic estimate for Eocene/Oligocene ASL fall (70 m) (ref. 17) (Fig. 3b). This calculated Eocene/Oligocene ice volume is similar to that estimated for the maximum extent Last Glacial Maximum on Antarctica (where most of the increase relative to today was achieved by grounding-line advance to near the shelf-slope break all around the continent)¹⁹.

If, in response to a lower latitudinal temperature gradient, we assume that the average isotopic composition of Oligocene Antarctic ice was less extreme than today's (-30‰), the implied global ice budget and ASL fall for the Eocene/Oligocene are correspondingly greater (~ 2.7 times, Fig. 3c), and impossibly large for Antarctica alone, given its combined continent and shelf area and the limits imposed on ice-sheet thickness by the strong dependence of ice flow on stress^{19,20}. These observations raise the possibility of contemporaneous Northern Hemisphere glaciation, consistent with evidence²¹ for early onset of North Atlantic Deep Water formation. In any case, our calculated ice volumes are so large that we conclude that Oi-1 must include some cooling component. This would imply that something acts to mask the cooling signal in the Mg/Ca records, possibly the effect of increasing sea-water pH and/or [CO₃²⁻] associated with CCD increase on Mg partitioning into foraminiferal calcite¹⁶.

Our data show that the Eocene/Oligocene CCD shift was synchronous with the development of major permanent Cenozoic Antarctic ice sheets. Calcite preservation on the sea floor depends on the saturation state of the deep ocean, the flux of organic carbon (C_{org}) to the sea floor and the ratio of this flux to the flux of CaCO₃ (refs 1, 22–24). Today, the increase in deep-sea [CO₃²⁻] associated with a 1-km deepening of the global lysocline yields a drawdown in atmospheric CO₂ of less than 25 μatm (refs 23, 24). Thus, CCD increase is unlikely to have triggered Antarctic glaciation. More probably, glaciation triggered the CCD shift. In fact, the CCD probably deepened to compensate for a reduction in the global ratio of CaCO₃ to C_{org} burial, as suggested by the $\delta^{13}\text{C}$ increase in benthic foraminiferal calcite (Fig. 1).

It seems improbable that the mechanism responsible was

enhanced global Ca²⁺ and CO₃²⁻ flux to the oceans, prompted by glacial weathering^{25–27}. Arguably, the most attractive way to explain the apparent teleconnection between the onset of Antarctic glaciation and CCD increase in the tropical Pacific is a shift of global CaCO₃ sedimentation from shelf to deep ocean basins^{25,28,29}. On 10–20-kyr timescales, the ocean is near saturation with respect to CaCO₃, such that any reduction in global shelf and reef carbonate sedimentation will promote increased ocean alkalinity, CCD increase and increased deep-ocean carbonate accumulation. Glacioeustatic sea-level fall associated with the growth of large Antarctic ice sheets would have reduced the size of the shelf carbonate reservoir, promoting higher deep-ocean [CO₃²⁻] and a deeper CCD. It would also have exposed widespread Upper Cretaceous and Lower Palaeogene limestones to erosion, thereby increasing global river inputs (and $\delta^{13}\text{C}$) of dissolved inorganic carbon and alkalinity, further increasing [CO₃²⁻], deepening the CCD and increasing seawater $\delta^{13}\text{C}$ (Fig. 1).

Another mechanism with the potential to contribute to Eocene/Oligocene CCD increase is an increase in global siliceous (at the expense of calcareous) plankton export production³⁰. The steady-state response to such a change in CaCO₃ export flux (assuming constant river inputs) is likely to be an increase in ocean pH, [CO₃²⁻] and the depth of the CCD. The power of these (and other) hypotheses to explain the link between Eocene/Oligocene Antarctic glaciation and so rapid, pronounced and permanent a CCD shift needs to be tested using a range of modelling techniques. □

Methods

Chronology

The basis for our new chronology is lithological proxy measurements (bulk density, colour reflectance and magnetic susceptibility) that were collected during ODP Leg 199 using the multi-sensor track (MST) core scanner⁸. We used the MST data, together with additional bulk $\delta^{13}\text{C}$, $\delta^{18}\text{O}$ and %CaCO₃ measurements from sites 1218 and 1219, to generate an aligned and stacked revised composite depth scale between sites 1218 and 1219. This allowed us to verify the completeness of sediment recovery, as well as the detailed cross-correlation of magnetic reversal and biostratigraphic events. Magnetochron C12n was identified in both sites 1218 and 1219. Chron C13n was recovered in site 1219 only but, by matching characteristic features between holes and sites, our detailed composite depth scale allowed us to constrain the position of C13n within site 1218 on a decimetre scale. Additional shipboard wire-line logging data from site 1218 confirm that the sediment recovered across the Eocene/Oligocene transition is representative of the *in situ* formation, including the double-step and the intervening 'plateau' across this transition in site 1218.

The MST proxy measurements, which in the Oligocene primarily reflect variations of %CaCO₃, allowed us to generate a high-resolution stacked record of %CaCO₃ from the MST data by regression with the bulk measurements. Throughout the Oligocene, these data show remarkably strong variations on orbital eccentricity timescales (~ 110 -kyr and 400-kyr periods), as suggested by the initial, low-resolution, shipboard timescale. The detailed chronology was generated by matching the benthic stable-isotope data from site 1218 with an astronomical template⁹. Our final age model was generated by first matching very clear ~ 400 -kyr and ~ 110 -kyr cycles in the stable-isotope data to the astronomical template, and fine-adjusting individual obliquity (~ 40 kyr) and climatic precession scale (~ 22 -kyr) cycles throughout the record. The ~ 400 -kyr eccentricity cycle is also present in the colour reflectance data, and was used to constrain the timescale in the upper part of the Eocene, where %CaCO₃ was very low. Amplitude variation of obliquity cycles in the $\delta^{18}\text{O}$ record agrees with a ~ 1.2 -Myr cycle in astronomically calculated Earth's obliquity, providing further constraints. Our chronology results in revised estimates for the ages of magnetic reversals between magnetochrons C12n and C13n and the age estimate for the Eocene/Oligocene boundary (Supplementary Information).

CaCO₃ data

Bulk sediment %CaCO₃ was measured in small (5–30 mg) discrete samples (3-cm spacing) using both standard high-precision coulometric methods and a new rapid-throughput continuous-flow mass-spectrometry technique. These data were also used to calibrate our proxy estimate 'calculated %CaCO₃', determined from a stacked record of whole-core sediment physical properties (gamma-ray attenuation and porosity estimate; magnetic susceptibility; lightness).

Stable-isotope data

We analysed $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ in well-preserved *Cibicides* benthic foraminifera that are believed to have lived on or just within the sediment–water interface. These were picked from a narrow size fraction (250–400 μm). Because of large inter-sample species abundance fluctuations we analysed a consistent mix of three species (two each of *C. havanensis*, *C. grimsdalei* and *C. subspiralatus*). Samples were cleaned ultrasonically and analysed at the Southampton Oceanography Centre using a Europa Geo 20-20 mass

spectrometer equipped with a automatic carbonate preparation system (CAPS). Results are reported relative to the Vienna Pee Dee Belemnite standard (VPDB). Standard external analytical precision, based on replicate analysis of in-house standards calibrated to NBS-19, is better than 0.1‰ for $\delta^{18}\text{O}$ and $\delta^{13}\text{C}$.

Received 1 September; accepted 25 October 2004; doi:10.1038/nature03135.

1. Broecker, W. S. & Peng, T.-H. The role of CaCO_3 compensation in the glacial to interglacial atmospheric CO_2 change. *Glob. Biogeochem. Cycles* **1**, 15–29 (1987).
2. Van Andel, T. H. Mesozoic/Cenozoic calcite compensation depth and the global distribution of calcareous sediments. *Earth Planet. Sci. Lett.* **26**, 187–194 (1975).
3. Kennett, J. P. & Shackleton, N. J. Oxygen isotopic evidence for the development of the psychrosphere 38 Myr ago. *Nature* **260**, 513–515 (1976).
4. Miller, K. G., Wright, J. D. & Fairbanks, R. G. Unlocking the ice house: Oligocene-Miocene oxygen isotopes, eustasy, and margin erosion. *J. Geophys. Res.* **96**, B4, 6829–6849 (1991).
5. Zachos, J. C., Quinn, T. M. & Salamy, K. A. High-resolution (10^4 years) deep-sea foraminiferal stable isotope records of the Eocene-Oligocene climate transition. *Palaeoceanography* **11**, 251–266 (1996).
6. Lear, C. H., Elderfield, H. & Wilson, P. A. Cenozoic deep-sea temperatures and global ice volumes from Mg/Ca in benthic foraminiferal calcite. *Science* **287**, 269–272 (2000).
7. DeConto, R. M. & Pollard, D. Rapid Cenozoic glaciation of Antarctica triggered by declining atmospheric CO_2 . *Nature* **421**, 245–249 (2003).
8. Shipboard Scientific Party 2002. Leg 199 summary. *Proc. ODP Init. Rep.* (eds Lyle, M. W., Wilson, P. A. & Jancsek, T. R.) **199**, 1–87 (2002).
9. Laskar, J. *et al.* Long term evolution and chaotic diffusion of the insolation quantities of Mars. *Icarus* **170**, 343–364 (2004).
10. Peterson, L. C., Backman, J. Late Cenozoic carbonate accumulation and the history of the carbonate compensation depth in the western equatorial Indian Ocean. *Proc. ODP Sci. Res.* (eds Duncan, R. A., Backman, J., Dunbar, R. B. & Peterson, L. C.) **115**, 467–489 (1990).
11. Salamy, K. A. & Zachos, J. C. Latest Eocene-Early Oligocene climate change and Southern Ocean fertility: inferences from sediment accumulation and stable isotope data. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **145**, 61–77 (1999).
12. Ehrmann, W. U. & Mackensen, A. Sedimentological evidence for the formation of an East Antarctic ice sheet in Eocene/Oligocene time. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **93**, 85–112 (1992).
13. Ivany, L. C., Patterson, W. P. & Lohmann, K. C. Cooler winters as a possible cause of mass extinction at the Eocene/Oligocene boundary. *Nature* **407**, 887–890 (2000).
14. Pälike, H., Laskar, J. & Shackleton, N. J. Geologic constraints on the chaotic diffusion of the solar system. *Geology* **32** (11), 929–932 doi: 10.1130/G20750 (2004).
15. Billups, K. & Schrag, D. P. Application of benthic foraminiferal Mg/Ca ratios to questions of Cenozoic climate change. *Earth Planet. Sci. Lett.* **209**, 181–195 (2003).
16. Lear, C. H., Rosenthal, Y., Coxall, H. K. & Wilson, P. A. Late Eocene to early Miocene ice-sheet dynamics and the global carbon cycle. *Palaeoceanography* **19**, doi:10.1029/2004PA001039 (2004).
17. Pekar, S. F., Christie-Blick, N., Kominz, M. A. & Miller, K. G. Calibration between eustatic estimates from backstripping and oxygen isotopic records for the Oligocene. *Geology* **30**, 903–906 (2002).
18. Lythe, M. B., Vaughan, D. G. & BEDMAP Consortium, BEDMAP: A new ice thickness and subglacial topographic model of Antarctica. *J. Geophys. Res.* **106**, B6, 11335–11351 (2001).
19. Huybrechts, P. Sea-level changes at the LGM from ice-dynamic reconstructions of the Greenland and Antarctic ice sheets during the glacial cycles. *Quat. Sci. Rev.* **21**, 203–231 (2002).
20. Hindmarsh, R. C. A. Time-scales and degrees of freedom operating in the evolution of continental ice-sheets. *Trans. R. Soc. Edinb. Earth Sci.* **81**, 371–384 (1990).
21. Davies, R., Cartwright, J., Pike, J. & Line, C. Early Oligocene initiation of North Atlantic deep water formation. *Nature* **410**, 917–920 (2001).
22. Archer, D. & Maier-Reimer, E. Effect of deep-sea sedimentary calcite preservation on atmospheric CO_2 concentration. *Nature* **367**, 260–263 (1994).
23. Sigman, D. M. & Boyle, E. A. Glacial/interglacial variations in atmospheric carbon dioxide. *Nature* **407**, 859–869 (2000).
24. Zeebe, R. E. & Westbroek, P. A simple model for the CaCO_3 saturation state of the ocean: The “Strangelove”, the “Neritan”, and the “Cretan” Ocean. *Geochem. Geophys. Geosyst.* **4**, 1104 (2003).
25. Kump, L. R. & Arthur, M. A. in *Tectonics Uplift and Climate Change* (ed. Ruddiman, W. F.) 399–426 (Plenum, New York, 1997).
26. Zachos, J. C., Opdyke, B. N., Quinn, T. M., Jones, C. E. & Halliday, A. N. Early Cenozoic glaciation, Antarctic weathering and seawater $^{87}\text{Sr}/^{86}\text{Sr}$: is there a link? *Chem. Geol.* **161**, 165–180 (1999).
27. Ravizza, G. & Peucker-Ehrenbrink, B. The marine $^{187}\text{Os}/^{188}\text{Os}$ record of the Eocene-Oligocene transition: the interplay of weathering and glaciation. *Earth Planet. Sci. Lett.* **210**, 151–165 (2003).
28. Berger, W. H. & Winterer, E. L. in *Plate Stratigraphy and the Fluctuating Carbonate line in Pelagic Sediments: On Land and Under the Sea* (eds Hsu, K. J. & Jenkyns, H. C.) 11–48 (Int. Assoc. Sedimentologists Spec. Publ. 1, Blackwell Science, Oxford, 1974).
29. Opdyke, B. N. & Wilkinson, B. H. Surface area control of shallow cratonic to deep marine carbonate accumulation. *Palaeoceanography* **3**, 685–703 (1989).
30. Harrison, K. G. Role of increased marine silica input on paleo- pCO_2 levels. *Palaeoceanography* **15**, 292–298 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the Shipboard Party of Ocean Drilling Program Leg 199 for assistance at sea and M. Bolshaw, M. Cooper and H. Birch for laboratory assistance. This work was supported by a NERC UK ODP grant to P.A.W., a Royal Commission for the Exhibition of 1851 fellowship awarded to H.K.C. and by Swedish Research Council (VR) funding to H.P. We thank W. Broecker, R. Hindmarsh, S. D'Hondt, A. Merico, Y. Rosenthal, R. Rickaby, J. Shepherd and T. Tyrrell for discussions and comments on an earlier draft and L. Kump for a constructive review.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.A.W. (paw1@soc.soton.ac.uk).

Similar response of labile and resistant soil organic matter pools to changes in temperature

Changming Fang¹, Pete Smith¹, John B. Moncrieff² & Jo U. Smith¹

¹School of Biological Sciences, University of Aberdeen, Aberdeen AB24 3UU, UK

²Ecology and Resource Management, School of GeoSciences, The University of Edinburgh, Edinburgh EH9 3JU, UK

Our understanding of the relationship between the decomposition of soil organic matter (SOM) and soil temperature affects our predictions of the impact of climate change on soil-stored carbon¹. One current opinion is that the decomposition of soil labile carbon is sensitive to temperature variation whereas resistant components are insensitive^{2–4}. The resistant carbon or organic matter in mineral soil is then assumed to be unresponsive to global warming^{2,4}. But the global pattern and magnitude of the predicted future soil carbon stock will mainly rely on the temperature sensitivity of these resistant carbon pools. To investigate this sensitivity, we have incubated soils under changing temperature. Here we report that SOM decomposition or soil basal respiration rate was significantly affected by changes in SOM components associated with soil depth, sampling method and incubation time. We find, however, that the temperature sensitivity for SOM decomposition was not affected, suggesting that the temperature sensitivity for resistant organic matter pools does not differ significantly from that of labile pools, and that both types of SOM will therefore respond similarly to global warming.

The temperature sensitivity of SOM decomposition, commonly referred to as Q_{10} , is critical for modelling changes in soil C stock^{3–6}. The assumption that the decomposition of old organic matter^{2–3} or organic C in mineral soil⁴ does not vary with temperature—that is, that the decomposition of labile C pools are sensitive, but resistant pools are insensitive, to temperature perturbations—suggests that higher losses of carbon will occur from soils in boreal and tundra regions in response to global warming. This is because these soils have the largest store of labile organic matter, and are predicted to experience the greatest rise in temperature⁷. Tropical soils may release less C than previously predicted⁴ owing to a large store of SOM in deep soil⁸ and the high proportion of resistant C pools in SOM. Soil warming experiments, an analogue for the effects of global warming on SOM decomposition⁹, suggest that the effect of warming on SOM decomposition may decline with time. The change in SOM composition associated with warming and the different temperature sensitivity of the C pools were assumed to be responsible for this decline^{10–11}. Despite the common assertion that SOM composition affects the temperature sensitivity of SOM decomposition, experimental or modelling evidence is yet to be presented. If the temperature sensitivity of SOM decomposition is not affected by SOM composition, predictions of climate change impacts on soil stored C will be greatly affected.

By definition, the temperature sensitivity of SOM decomposition is the change in SOM decomposition rate with temperature under otherwise constant conditions⁵. At present, this concept is often confused with concepts of SOM turnover^{4,12–13} or SOM dynamics^{2–4} under different environmental conditions with accompanying different temperatures. Temperature sensitivity of SOM decomposition (or Q_{10}) estimated by incubating soils at different but constant temperatures^{14–16} or by radiocarbon accumulation in undisturbed soils¹³ is confounded by many factors other than temperature.

We incubated soil samples under changing temperature to

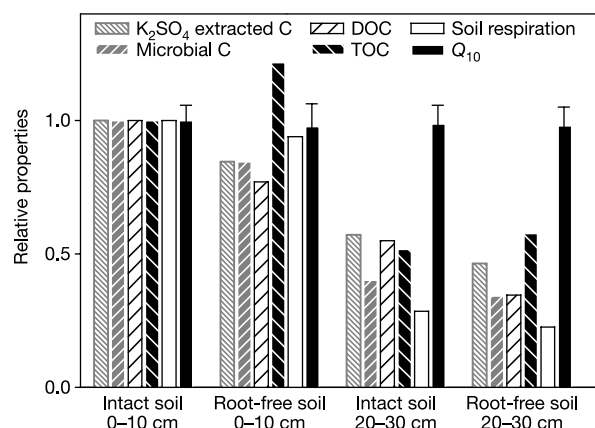


Figure 1 Soil carbon components, respiration rate and associated Q_{10} values with respect to soil depth and sampling method (four replicates for each sample). Respiration rate was an average of data measured at 20 °C in days 3 and 5. The Q_{10} value was estimated with soil respiration rates under changing temperature for the period of days 3–10. All values are normalized against that of surface root-free sample. Soil respiration rate is significantly related to concentrations of C pools owing to soil depth and sampling method, but Q_{10} does not change with respiration rate or C concentrations. Error bars indicate standard deviation. DOC, dissolved organic carbon; TOC, total organic carbon.

investigate the influence of SOM composition on the temperature dependence of SOM decomposition. Figure 1 shows that soil C contents for both labile components (water-dissolved organic carbon (DOC), microbial carbon (C_{mic}) and K_2SO_4 -extracted carbon (C_{KSO})) and the total organic carbon (TOC), are significantly lower in the subsoil (20–30 cm) than in the surface soil (0–10 cm). The ratio of DOC:TOC and C_{KSO} :TOC declined significantly with soil depth ($F = 28.5$ and 36.1 , respectively, $P < 0.0001$), but C_{mic} :TOC was not significantly affected by depth ($F = 1.9$, $P < 0.2$). After the initial flush of CO_2 emission, soil basal respiration rate at 20 °C was (mean $\pm$ s.e.m.) $6.67 \pm 0.46 \mu g CO_2$ per g dry soil per h for root-free samples in the 0–10 cm layer, but only $1.92 \pm 0.20 \mu g CO_2$ per g dry soil per h for the 20–30 cm layer. Corresponding values were 6.27 ± 0.66 and 1.47 ± 0.16 for intact samples. Over a period up to 88 days, the subsoil respired only $\sim 0.29 \pm 0.13$ of the CO_2 respired in the surface soil. These results indicate that soil basal respiration rate is closely related to variations in C pools occurring at different soil depths.

Q_{10} values for individual soil samples varied in the range 1.97–2.21 during the early stage of incubation (up to day 10). No significant correlation was found between Q_{10} and the rate of basal respiration. Relationships in Fig. 1 between respiration rate, Q_{10} value and SOM pools reflect the long-term acclimation of the microbial community to the environment (such as temperature, moisture and O_2) associated with soil depths.

During the incubation, there was a significant decline in the labile components (Fig. 2c, d, f). After 108 days incubation, DOC was 0.73 ± 0.14 and C_{KSO} was 0.62 ± 0.065 of initial values when averaged over all samples. The greatest variation following incubation was observed in C_{mic} . The average C_{mic} at day 42 was only 0.43 ± 0.13 of the initial content, and less than 0.10 ± 0.0057 after 108 days incubation. Changes in the average TOC during incubation were not significant (Fig. 2b). At the end of the incubation, average TOC was 0.94 ± 0.19 of the initial content. Soil respiration rate consistently declined with time (Fig. 2e). The association between respiration rate and C_{mic} during the incubation suggests that the variation in microbial biomass may be a major cause of the temporal changes in soil respiration.

The response of soil basal respiration to temperature was not

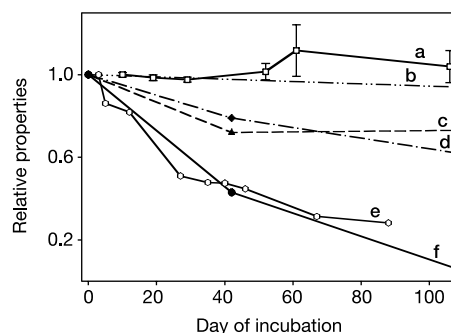


Figure 2 Variations in respiration rate and soil carbon pools with increasing incubation time. Values are averages of all four samples, and normalized by initial values. **a**, Q_{10} value; **b**, TOC; **c**, DOC; **d**, K_2SO_4 -extracted C; **e**, respiration rate at 20 °C; and **f**, microbial biomass C. Error bars are standard deviation. Respiration rate declined rapidly owing to the depletion of labile components (DOC, C_{KSO} and C_{mic}), but the Q_{10} value of soil respiration remained unchanged.

affected by the depletion of labile C during the incubation. Q_{10} values averaged for all samples were in the range 2.01–2.30 for the whole incubation period (Fig. 2a). There is no significant change in Q_{10} for soil basal respiration with incubation time, despite the fact that Q_{10} was more variable during the later stages of incubation. As time progressed, the resistant C component contributed a greater portion of the total soil basal respiration owing to the depletion of labile C pools (Supplementary Fig. 2). The Q_{10} value for soil basal respiration should gradually decrease if resistant C is significantly different from labile pools and insensitive to temperature variation (Supplementary Fig. 3). A constant Q_{10} for soil basal respiration suggests that the temperature dependence of resistant C is not significantly different from that for labile pools.

In most incubation experiments, soil samples have been separately incubated at different but constant temperatures^{12,14,15}. Three different methods have been used to estimate SOM decomposition and its temperature sensitivity: the total mass loss^{3,17}, the time required for a given percentage of mass loss¹⁷, and the soil respiration rate^{14,18}. A decline in soil respiration rate was commonly observed as incubation times increased^{14,17,19,20}. This decline is expected to be greater at higher than at lower temperatures because of the greater depletion and degradation of C pools²¹. Temperature sensitivity is likely to be underestimated if turnover rate is derived from studies of total mass loss for a given time period or from respiration rates at different constant temperatures, owing to the higher decline in C turnover rate at higher temperature. If Q_{10} is estimated using the time required for a given percentage of mass loss, the value will be overestimated. In this case, temporal effects on estimated C turnover rate are more pronounced at lower temperatures than at higher temperatures. Data of total mass loss from soil incubations longer than one year were used to support the opinion that decomposition rates of organic matter in mineral soil do not vary with temperature⁴. Estimated C turnover rates from a long-term incubation will be significantly different from those occurring in the field, owing to the quick decline in soil microbial biomass and respiration rate during incubation. In such experiments, the temperature sensitivity of SOM decomposition may have been seriously biased or underestimated because respiration rates at all temperatures are close to zero at the later stage of incubation.

In soil warming experiments, the observed decline of warming effects on SOM decomposition with time¹¹ does not necessarily mean that the decomposition of resistant C is less sensitive to elevated temperature than the labile component. Provided that the increase in net primary production (NPP) due to warming is less

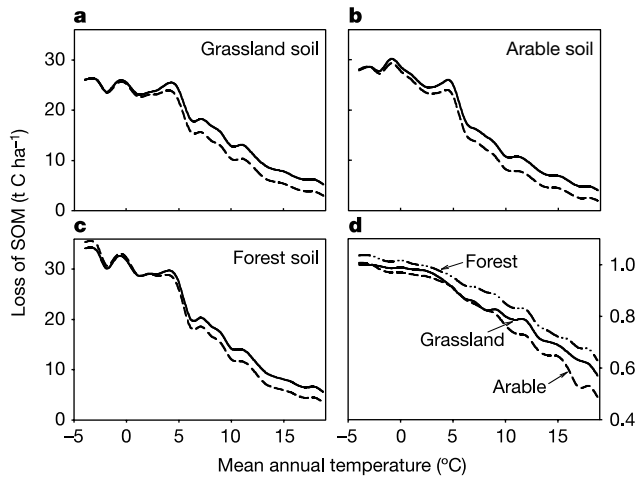


Figure 3 Changes in soil C by 2100 for European soils. The baseline (solid lines in **a–c**) was from the original Roth-C model²³ projection ($Q_{10} = 2.98$ at 10°C for all C pools). The temperature sensitivity of humus was changed to 80% of the original value ($Q_{10} = 2.58$ at 10°C , dashed lines in **a–c**). The loss of soil C is an average of all grid cells ($21,976$ cells at $10' \times 10'$ resolution) according to present MAT. The percentage of net soil C loss with modified $Q_{10} = 2.58$ for humus is relative to baseline decreases with MAT gradient (**d**). SOM, soil organic matter.

than the increase in SOM decomposition rate, a decline in warming effect on SOM decomposition is always expected. In the long-term, the microbial community may become acclimated to warming with changed activities. The contribution of this acclimation is not yet clear.

For long-term climate change, the response of the resistant pool of SOM plays a critical role in regulating soil C stocks. Given the predicted climate change in Europe in the next century, the greatest loss of SOM is expected in soils where the present mean annual temperature (MAT) is less than 4°C , and this net release of SOM will gradually decrease with MAT gradient (Fig. 3a–c). (This predicted climate change is climate forcing according to the implementation by the Hadley Centre Climate Model (HadCM3) of the Intergovernmental Panel on Climate Change (IPCC) A1FI (world market–fossil fuel intensive) emission scenario²².) With a moderate change in the temperature sensitivity of the resistant C pool (humus pool of the Rothamsted Carbon Model²³ only), from $Q_{10} = 2.98$ to about 2.58 (at 10°C), sensitivity induced change will significantly reduce the net SOM release in temperate soils (present MAT $> 4^{\circ}\text{C}$). By 2100, the reduction in SOM loss could be up to 46% in arable soil, 37% in grassland, and 32% in forest for regions where the present MAT is greater than 15°C (Fig. 3d). At the global scale, this reduction will be large enough to change our prediction of the magnitude and spatial pattern of SOM stocks in the future. Our study does not support the opinion that resistant C pools are significantly less responsive to temperature variation than labile C pools. □

Methods

Soil samples (intact and root-free) were collected from a middle-aged plantation of Sitka spruce (*Picea sitchensis*) in Scotland ($56^{\circ}37' \text{N}$, $3^{\circ}48' \text{W}$). Mineral soils were collected from four locations in the site at depths of 0–10, 20–30 cm. Root-free samples were made by sieving soil through a 2 mm mesh to remove plant detritus, root and gravel. For each depth, approximately 600–800 g soil was taken and packed into a chamber to the original bulk density. Intact soil samples ($\sim 10 \times 10 \text{ cm}$) were taken next to each root-free sample, following the method of ref. 5. Soil samples were analysed to determine TOC^{24} , DOC^{25} and C_{KSO} . C_{mic} was determined by fumigation extraction²⁶. Samples (16 in total) were incubated in the laboratory using a programmable water bath (developed in The University of Edinburgh, UK). Temperature was changed commonly between 4 and 44°C (continuously increased from the lowest to the highest with a step of 4°C and then decreased, reaching a new temperature within two hours). Each temperature was held for

about 9 h. Before and after each round of temperature change, soils were kept at 20°C for a few days. Soil moisture contents were monitored and adjusted accordingly by adding water at the surface of the soil sample, and fresh air was continuously passed through each chamber during the incubation. Respired CO_2 was measured with an infrared gas analyser in differential mode, logged every second for 7 min for each chamber, but only the average over the last four minutes was used. The 16 chambers were measured sequentially, and four rounds of measurement were made before changing to another temperature. During each round of temperature change, the mean respiration rate at a given temperature was an average of values measured at that temperature when the temperature was increasing and decreasing (Supplementary Fig. 1). Mean respiration rates at different temperatures were fitted with an exponential model² ($R = a \exp[\ln(Q_{10}(T/10))]$) to calculate the Q_{10} value. More information about data analysis is included in the Supplementary Methods, which also explain how we assessed contributions of the resistant C pool to the total SOM decomposition and its Q_{10} .

Received 30 July; accepted 22 October 2004; doi:10.1038/nature03138.

- Lenton, T. M. & Huntingford, C. Global terrestrial carbon storage and uncertainties in its temperature sensitivity examined with a simple model. *Glob. Change Biol.* **9**, 1333–1352 (2003).
- Liski, J., Ilvesniemi, H., Mäkelä, A. & Westman, C. J. CO_2 emissions from soil in response to climatic warming are overestimated—The decomposition of old soil organic matter is tolerant of temperature. *Ambio* **28**, 171–174 (1999).
- Thornley, J. H. M. & Cannell, M. G. R. Soil carbon storage response to temperature: a hypothesis. *Ann. Bot.* **87**, 591–598 (2001).
- Giardina, C. P. & Ryan, M. G. Evidence that decomposition rates of organic carbon in mineral soil do not vary with temperature. *Nature* **404**, 858–861 (2000).
- Fang, C. & Moncrieff, J. B. The dependence of soil CO_2 efflux on temperature. *Soil Biol. Biochem.* **33**, 155–165 (2001).
- Sanderman, J., Amundson, R. G. & Baldocchi, D. D. Application of eddy covariance measurements to the temperature dependence of soil organic matter mean residence time. *Glob. Biogeochem. Cycles* **17**, doi:10.1029/2001GB001833 (2003).
- Schlesinger, W. H. & Andrews, J. A. Soil respiration and the global carbon cycle. *Biogeochemistry* **48**, 7–20 (2000).
- Jobbágy, E. G. & Jackson, R. B. The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecol. Appl.* **10**, 423–436 (2000).
- Rustad, L. E. *et al.* A meta-analysis of the response of soil respiration, net nitrogen mineralization, and aboveground plant growth to experimental ecosystem warming. *Oecologia* **126**, 543–562 (2001).
- Peterjohn, W. T., Melillo, J. M. & Bowles, S. T. Soil warming and trace gas fluxes: experimental design and preliminary flux results. *Oecologia* **93**, 18–24 (1993).
- Peterjohn, W. T. *et al.* Response of trace gas fluxes and N availability to experimentally elevated soil temperature. *Ecol. Appl.* **4**, 617–625 (1994).
- Dalias, P., Anderson, J. M., Bottner, P. & Couëteux, M.-M. Temperature responses of carbon mineralization in conifer forest soils from different regional climates incubated under standard laboratory conditions. *Glob. Change Biol.* **6**, 181–192 (2001).
- Trumbore, S. E., Chadwick, O. A. & Amundson, R. Rapid exchange between soil carbon and atmospheric carbon dioxide driven by temperature change. *Science* **272**, 393–396 (1996).
- Winkler, J. P., Cherry, R. S. & Schlesinger, W. H. The Q_{10} relationship of microbial respiration in a temperate forest soil. *Soil Biol. Biochem.* **28**, 1067–1072 (1996).
- MacDonald, N. W., Zak, D. R. & Pregitzer, K. S. Temperature effects on kinetics of microbial respiration and net nitrogen and sulfur mineralization. *Soil Sci. Soc. Am. J.* **59**, 233–240 (1995).
- Ross, D. J. & Tate, K. R. Microbial C and N, and respiratory activity, in litter and soil of a southern beech (*Nothofagus*) forest: distribution and properties. *Soil Biol. Biochem.* **25**, 477–483 (1994).
- Reichstein, M., Bednorz, F., Broll, G. & Kätterer, T. Temperature dependence of carbon mineralisation: conclusions from a long-term incubation of subalpine soil samples. *Soil Biol. Biochem.* **32**, 947–958 (2000).
- Fierer, N., Allen, A. S., Schimel, J. P. & Holden, P. A. Controls on microbial CO_2 production: a comparison of surface and subsurface soil horizon. *Glob. Change Biol.* **9**, 1322–1332 (2003).
- Lovell, R. D. & Jarvis, S. C. Soil microbial biomass and activity in soil from different grassland management treatments stored under controlled conditions. *Soil Biol. Biochem.* **30**, 2077–2085 (1998).
- Lomander, A., Kätterer, T. & Andrén, O. Carbon dioxide evolution from top- and subsoil as affected by moisture and constant and fluctuating temperature. *Soil Biol. Biochem.* **30**, 2017–2022 (1998).
- Grisi, B., Grace, C., Brookes, P. C., Benedetti, A. & Dell'abate, M. T. Temperature effects on organic matter and microbial biomass dynamics in temperate and tropical soils. *Soil Biol. Biochem.* **30**, 1309–1315 (1998).
- IPCC. *Special Report on Emissions Scenarios* (Cambridge Univ. Press, Cambridge, UK, 2000).
- Coleman, K. & Jenkinson, D. S. in *Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets* (eds Powlson, D. S., Smith, P. & Smith, J. U.) 237–246 (NATO ASI Series I Vol. 38, Springer, Heidelberg, 1996).
- Allen, S. E., Grimshaw, H. M., Parkinson, J. A. & Quarmby, C. *Chemical Analysis of Ecological Materials* 137–139 (Blackwell Scientific, Oxford, 1974).
- Martin-Olmedo, P. & Rees, R. M. Short-term N availability in response to dissolved organic-carbon from poultry manure, alone or in combination with cellulose. *Biol. Fert. Soils* **29**, 386–393 (1999).
- Öhlinger, R. in *Methods in Soil Biology* (eds Schinner, F. *et al.*) 56–58 (Springer, Berlin, 1995).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Wattenbarch and C. Zhang for assistance with the modelling. The pan-European modelling used data sets arising from the EU-funded ATEAM project.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.F. (c.fang@abdn.ac.uk).

Evolution driven by differential dispersal within a wild bird population

Dany Garant¹, Loeske E.B. Kruuk², Teddy A. Wilkin¹, Robin H. McCleery¹ & Ben C. Sheldon¹

¹Edward Grey Institute, Department of Zoology, University of Oxford, Oxford OX1 3PS, UK

²Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK

Evolutionary theory predicts that local population divergence will depend on the balance between the diversifying effect of selection and the homogenizing effect of gene flow^{1–3}. However, spatial variation in the expression of genetic variation will also generate differential evolutionary responses. Furthermore, if dispersal is non-random it may actually reinforce, rather than counteract, evolutionary differentiation. Here we document the evolution of differences in body mass within a population of great tits, *Parus major*, inhabiting a single continuous woodland, over a 36-year period. We show that genetic variance for nestling body mass is spatially variable, that this generates different potential responses to selection, and that this diversifying effect is reinforced by non-random dispersal. Matching the patterns of variation, selection and evolution with population ecological data, we argue that the small-scale differentiation is driven by density-related differences in habitat quality affecting settlement decisions. Our data show that when gene flow is not homogeneous, evolutionary differentiation can be rapid and can occur over surprisingly small spatial scales. Our findings have important implications for questions of the scale of adaptation and speciation, and challenge the usual treatment of dispersal as a force opposing evolutionary differentiation.

Theoretically, adaptive population divergence depends on the balance between diversifying natural selection and homogenizing gene flow, with one force opposing the other^{4,5}. Variability in habitat quality can generate different selection pressures and can thus result in phenotypic divergence⁶ and potentially even sympatric speciation^{7,8}, as will differences in the genetic architecture of phenotypic traits and hence any evolutionary response to natural selection. So far, however, there have been very few documented examples of these processes from wild populations in natural environments. In addition, whereas most models of evolution under spatially variable selection assume that dispersal is random, this need not be the case, especially when there is an interaction between phenotype and habitat quality^{9,10} such as might result from social dominance of large over small individuals¹¹. Theoretically, differential dispersal could then act to reinforce differentiation generated by the response to diversifying selection. However, it is unclear whether these processes do in fact operate in this way in natural populations.

We examined spatial variation in the selection and evolution of fledging mass as documented during a long-term study of a population of individually marked great tits (*Parus major*) at Wytham, Oxfordshire, UK (51°47' N, 1°20' W), from 1965 to 2000. Fledging mass is an excellent model trait because it is subject to positive directional selection^{12,13} (see Fig. 1), can be easily measured on large samples of relatives, and shows moderate heritability in this and other species of birds^{13–15}. In this population, great tits show natal dispersal behaviour only, with very little breeding dispersal outside the area in which they settle as first-time breeders¹⁶, which facilitates the assessment of dispersal patterns of individuals.

Wytham is a heterogeneous woodland, arbitrarily divided into

nine sectors that show different habitat characteristics¹⁷. A close examination of phenotypic and genotypic (estimated by individual breeding values¹⁸) trends for fledging mass over our 36-year study period revealed highly significant spatio-temporal heterogeneity among the eight sectors of the woodland studied for the whole period (sector-by-year interaction: phenotype, $\chi^2_{(7)} = 40.79$, $P < 0.001$; genotype, $\chi^2_{(7)} = 116.78$, $P < 0.001$). More detailed analyses reveal strikingly different patterns depending on sector, with a strong spatial organization of the pattern observed among sectors (Fig. 2a). Specifically, sectors in the eastern part of the wood (Marley Wood and Marley Plantation, henceforth referred to as east; see Fig. 2a) showed the same general trend, whereas sectors in the northern part of the wood (Extra, Great Wood and Common Piece, henceforth called north; see Fig. 2a) showed a different tendency. The remaining sectors, located in the central part of the wood, showed an intermediate response (Fig. 2a) and will not be discussed in the remaining analyses. These sectors receive substantial immigration from both east and north (see Supplementary Information) and are therefore expected to behave in an intermediate fashion. An analysis based on individual nest boxes (which had the same location for the entire study period), without any reference to sector, confirmed the general pattern of clustering into east and north blocks (Fig. 2b). We therefore sought to explain the differences between the areas (east and north) with the most contrasting patterns.

Birds in the eastern part of the wood showed a decrease in mean phenotype (Fig. 3a) and breeding value (Fig. 3b) over time. In contrast, birds from the northern part of the wood showed no temporal trend in their mean phenotype (Fig. 3a) but a highly significant increase in their breeding value over time (Fig. 3b). Consequently, the two parts of the population showed significantly different patterns of divergence in both phenotype (difference between slopes = 0.019 ± 0.0081 , $t_{67} = 2.35$, $P = 0.022$; estimates are presented throughout with their standard errors, s.e.m.) and underlying genotype (difference between slopes = 0.0089 ± 0.0012 , $t_{67} = 7.32$, $P < 0.001$). Linear mixed modelling, controlling for non-independence of chicks from a brood, confirmed these significant differences between areas over time (phenotypes,

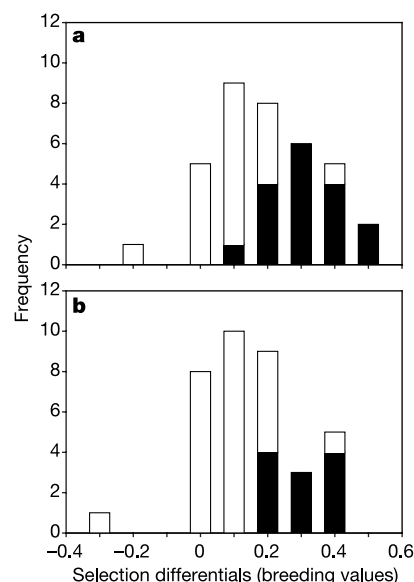


Figure 1 Distribution of standardized directional selection differentials for great tits for each year from 1965 to 2000 in Wytham¹³. **a**, Fledging mass phenotype; **b**, fledging mass breeding values. The shaded proportions of each bars indicate selection differentials significantly ($P < 0.05$) different from zero in individual years.

0.0185 ± 0.0031 , $\chi^2_{(1)} = 34.90$, $P < 0.001$; breeding values, 0.0085 ± 0.0008 , $\chi^2_{(1)} = 102.27$, $P < 0.001$; see also Supplementary Information). As a result, the population in the eastern part of the wood, which was composed of birds with the highest average fledging mass phenotype and breeding value at the beginning of the study period, is now composed of the smallest birds.

Different patterns of evolution might be caused by different patterns of selection, differing responses to selection, differences in dispersal, or a combination of these processes. Nestling fledging mass was under significant directional selection for survival in both areas (selection intensity: east, 0.140 ± 0.032 , $\chi^2_{(1)} = 27.91$, $P < 0.001$; north, 0.179 ± 0.023 , $\chi^2_{(1)} = 85.42$, $P < 0.001$) and

the same was true for the breeding values for this character (east, 0.077 ± 0.032 , $\chi^2_{(1)} = 8.48$, $P = 0.004$; north, 0.128 ± 0.023 , $\chi^2_{(1)} = 36.46$, $P < 0.001$). The intensity of directional selection did not differ between areas (phenotype, $t_{24917} = 0.99$, $P = 0.322$; breeding value, $t_{24917} = 1.30$, $P = 0.194$).

The response to selection depends on the amount of additive genetic variance expressed; a given character can express differing quantities of genetic variance in different environments¹⁹. In birds, size-related heritabilities are generally higher under favourable environmental conditions than under poorer conditions, in which a smaller component of additive genetic variance is commonly observed¹⁹. In our study population we found significantly greater

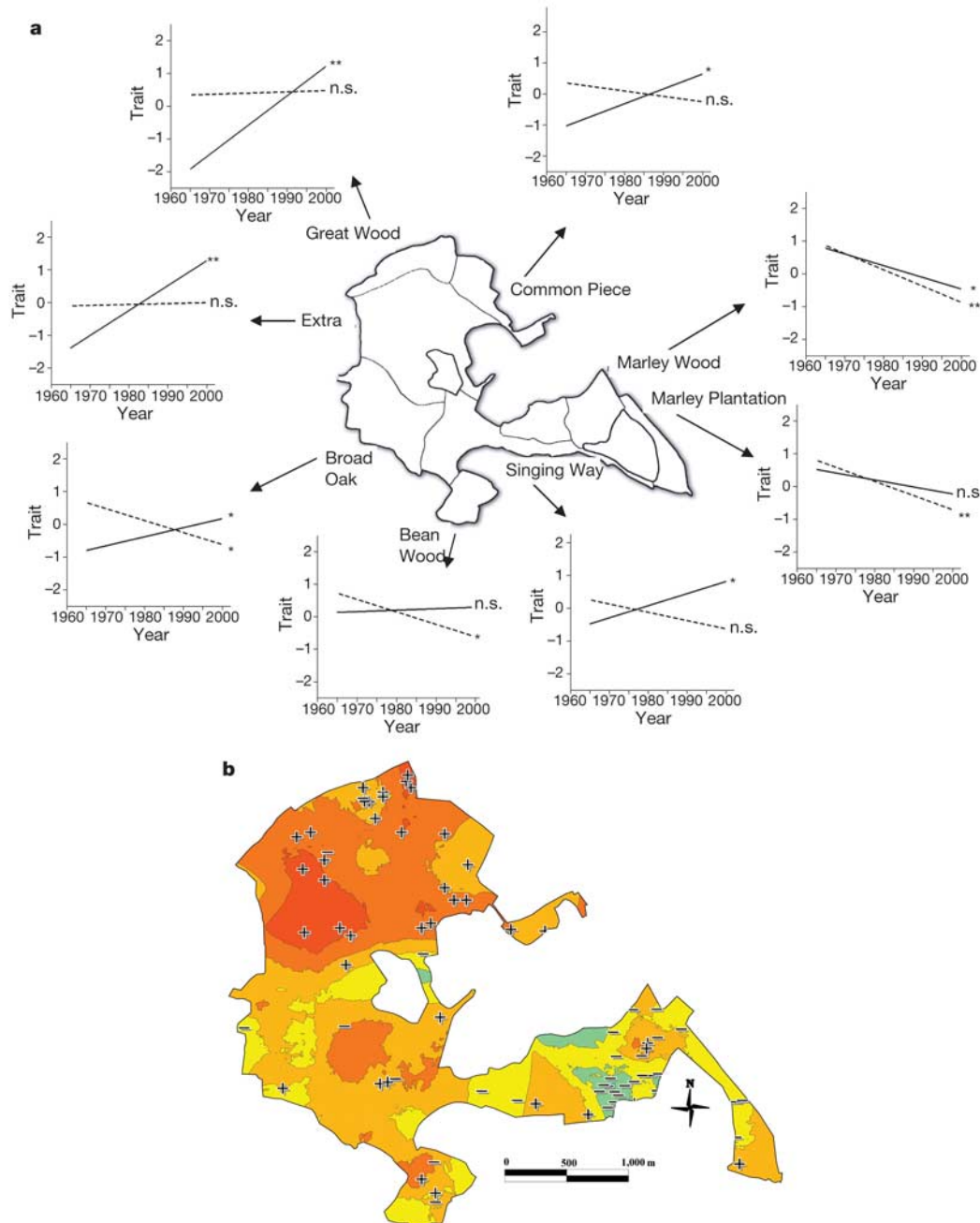


Figure 2 Within-population differences in great tit fledging mass trends from 1965 to 2000. **a**, Temporal patterns of variation for each sector of Wytham at phenotypic (dashed lines) and genotypic (solid lines) levels. Significance: n.s., not significant; one asterisk, $P < 0.05$; two asterisks, $P < 0.01$, referring to trends over time. **b**, Temporal changes in the genotypic component (estimated breeding value) of fledging mass

analysed at the level of individual nest boxes. Colours indicate local trends over time: red, increase; green, decrease. Plus signs, individual nest boxes that showed a significant increase; minus signs, individual nest boxes that showed a significant decrease.

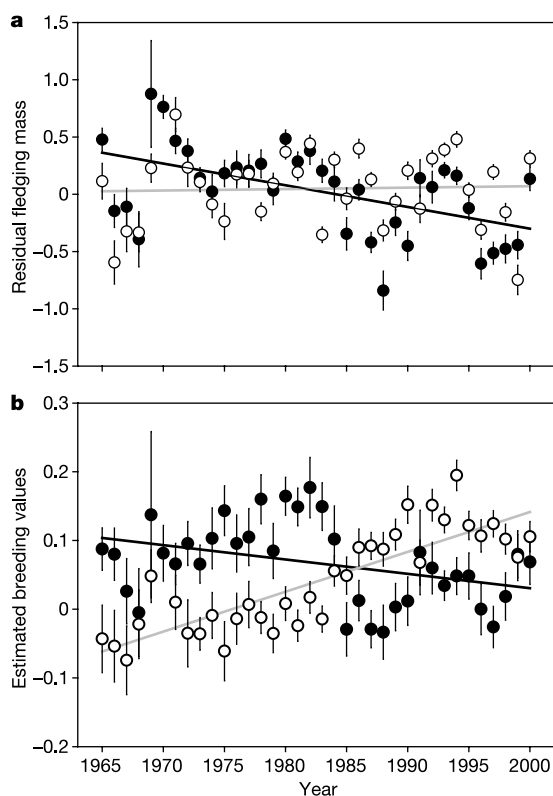


Figure 3 Temporal trends in fledging mass of great tit nestlings born within the eastern and northern parts of Wytham from 1965 to 2000. Mean residual fledging mass (**a**) and breeding values (**b**) (values are means $\pm$ 95% confidence intervals). Nestlings born in the north (grey line and open circles) showed no change in their phenotype ($b = 0.00024 \pm 0.00611$, $t_{33} = 0.04$, $P = 0.969$) but a highly significant increase in their breeding value over the study period ($b = 0.00633 \pm 0.00078$, $t_{33} = 8.11$, $P < 0.001$; weighted by annual sample size). Birds born in the east (black line, filled circles) showed a significant decrease at both levels (phenotype, $b = -0.01870 \pm 0.00514$ g yr $^{-1}$, $t_{34} = -3.64$, $P < 0.001$; breeding values, $b = -0.00255 \pm 0.00094$, $t_{34} = -2.73$, $P = 0.010$). The difference between the slopes is significant for both measures (see the text).

additive genetic variance in the north than in the east, which translated into a significantly higher heritability of fledging mass in this area (see Table 1). As a consequence, the expected response to selection was greater in the northern part of the wood than in the east. Using the simplest approach, in which the response to selection (R) is the product of selection and heritability of a character, scaled by generation time ($R = h^2 S / \text{number of generations}$), with a generation time of 1.97 years in the east and 1.83 years in the north, the expected response in the east ($+0.0141$ g yr $^{-1}$) is less than half that in the north ($+0.0287$ g yr $^{-1}$). Thus, this difference between areas in the expected response to selection, generated by variable levels of additive genetic variance, could contribute to the pattern of phenotypic differentiation observed between areas.

Although expected responses to selection thus differed depending

on the areas under study, these differences would seem likely to be swamped by gene flow, given that, on average, 62% of the breeding birds in a given area are born outside that area (see Supplementary Information), and they would not in any case explain the decline observed in the east. However, we also found marked evidence for differential dispersal, both within the study population and in terms of birds immigrating to the study area, both of which act to reinforce the different expected responses to selection documented above. Specifically, birds born in the central part of the wood showed non-random natal dispersal patterns (Fig. 4a), with birds emigrating to the north showing no change in mean phenotype over time (adult body mass from 1978 to 2000, $b = 0.0183 \pm 0.0116$, $t_{21} = 1.58$, $P = 0.130$), whereas birds emigrating to the east showed a significant decline in their body mass ($b = -0.0494 \pm 0.0123$, $t_{18} = -4.01$, $P < 0.001$); the difference between areas was significant (see Fig. 4a). This pattern of non-random dispersal was also present when analysing the phenotypic trends for birds immigrating to the wood but born outside Wytham (see Fig. 4b) (adult body mass from 1974 to 2000: east, $b = -0.0507 \pm 0.0084$, $t_{22} = -6.08$, $P < 0.001$; north, $b = -0.0161 \pm 0.0062$, $t_{23} = -2.58$, $P = 0.017$); the difference between areas was again significant (Fig. 4b). This pattern is unlikely to have been caused by the influence of the environment on adult mass of birds after settlement, because the change from fledging to adult mass (for birds from the central area, measured as nestlings) did not differ between areas over time (see Fig. 4c). The pattern therefore reflects differential distribution according to mass before settlement. A within-family comparison, restricted to broods in which one or more offspring immigrated to each of the north and east areas (267 nestlings from 114 families), confirmed the dependence of dispersal on phenotype (Fig. 4d). Within families, heavier offspring showed an increasing tendency over time to settle in the north of the wood. Settlement of birds is therefore markedly non-random, is dependent on phenotype, and contributes to the within-population differentiation.

The collection of long-term data on population dynamics from this study population indicates the ecological mechanism responsible for local differentiation. Specifically, our results indicate that the north has become the preferred habitat over time. In a previous study, we showed that density-dependent processes are important determinants of the uncoupling between genotypic and phenotypic trends in fledging mass over time in this population¹³. Higher density had a detrimental effect on mean fledging mass, leading to a larger negative environmental deviation of the phenotype from its genotypic value at the population level (that is, birds were lighter than expected)¹³. However, although the number of nest boxes and their locations were kept constant throughout the study period, breeding density varied both spatially and temporally within the study population (Fig. 5) and it is therefore a potentially important factor at a local scale within this population. Average local breeding density in the east was more than twice that in the north (mean occupied nest-box Theissen polygon area $\pm$ SD: east = $4,253 \pm 1,225$ m 2 , north = $10,097 \pm 2,187$ m 2 , paired t -test comparing annual values, $t_{28} = 13.71$, $P < 0.001$; see Fig. 5), indicating higher pressures on the habitat in the east than in the north. However, overall population density increased over the study period¹³, and we found a significant decrease in mean polygon

Table 1 Components of phenotypic variance in great tit fledging mass (g) and resulting heritability in contrasting areas of Wytham Woods

	N (nestlings)	Trait mean (s.d.)	Additive genetic variance (s.e.m.)	Environmental variance (s.e.m.)	Phenotypic variance	h^2 (s.e.m.)
East	9,237	18.7 (1.5)	0.220 (0.036)	0.884 (0.066)	1.104	0.199 (0.031)
North	15,682	18.8 (1.4)	0.315 (0.027)	0.756 (0.046)	1.071	0.294 (0.024)
P			0.035	0.113		0.015

Significance (P) of the differences between areas is based on a t -test; environmental variance is defined as the sum of brood, year and residual variance.

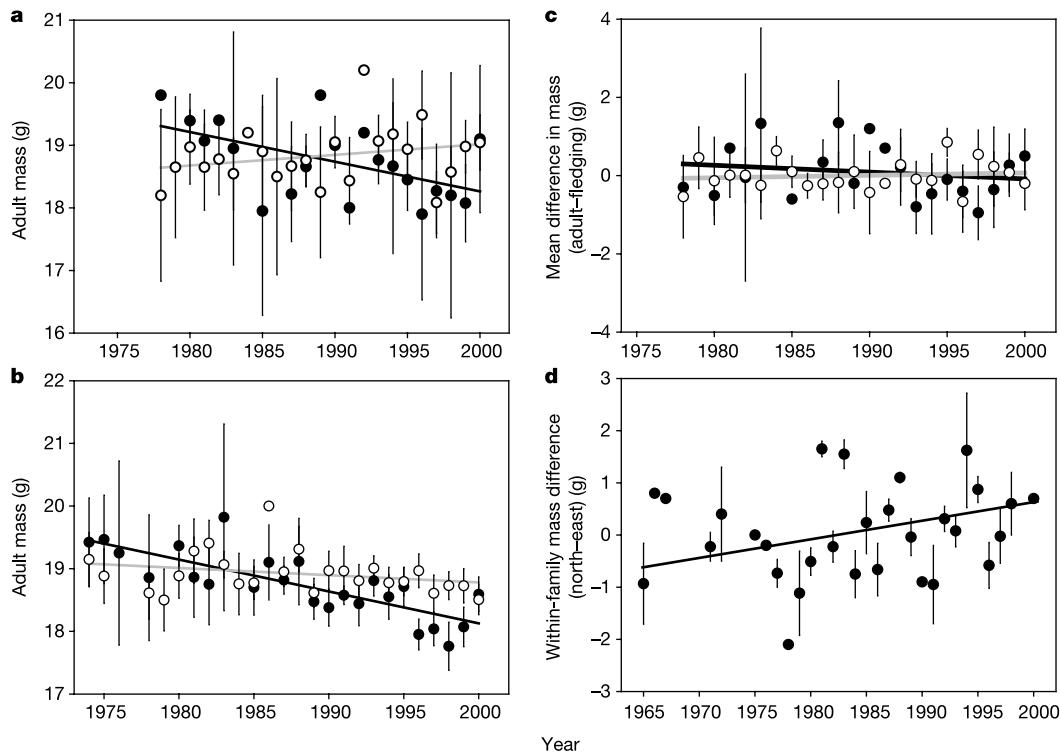


Figure 4 Mean adult mass of great tits immigrating to the eastern and northern parts of Wytham. The distribution of phenotypes with respect to time and area is non-random, because birds settling in the east (black line, filled circles) showed a much greater change in mass over time than birds settling in the north (grey line, open circles) (values are means $\pm$ 95% confidence intervals). **a, b**, The difference between the slopes is significant for both birds originating from the central part of the wood (0.0676 ± 0.0176 , $t_{39} = 3.84$, $P < 0.001$) (**a**) and birds originating from outside the woodland (0.0347 ± 0.0104 , $t_{45} = 3.33$, $P = 0.002$) (**b**). **c**, Difference between adult and fledging masses for central emigrants settling in the east (black line, filled circles) or the

north (grey line, open circles) of Wytham. The difference between areas over time is not significant (0.0093 ± 0.0233 , $t_{39} = 0.40$, $P = 0.692$), indicating that the area of settlement does not influence the adult mass difference observed in **a, d**. Mean difference in fledging mass of great tits from the same family settling into the east and north parts of Wytham. The distribution of phenotypes changed over time, with the difference in mass between individuals settling in the two areas changing from being negative to positive over the study period. The difference between areas over time is significant (linear mixed model controlling for brood identity: 0.0311 ± 0.0127 , $\chi^2_{(1)} = 6.01$, $P = 0.014$).

area over time in the north ($-156.5 \pm 35.5 \text{ m}^2 \text{ yr}^{-1}$, $t_{27} = -4.41$, $P < 0.001$), with no significant change in the east ($-34.4 \pm 24.6 \text{ m}^2 \text{ yr}^{-1}$, $t_{28} = -1.40$, $P = 0.173$; significant difference between the two trends; $b = 122.2 \pm 42.8 \text{ m}^2 \text{ yr}^{-1}$, $t_{55} = 2.85$, $P = 0.006$). The temporal difference between the two areas might indicate an effect of preferential settlement in higher-quality (lower-density) habitat, a suggestion reinforced by the phenotype-dependent dispersal, which is consistent with larger, heavier, individuals being able to settle in preferred habitats. We tested whether habitat preference was linked with an increase in fitness. We first found that the survival of chicks to recruitment did not differ significantly between areas throughout the study period (difference = 0.0029 ± 0.0040 , $\chi^2_{(1)} = 0.52$, $P = 0.470$), but that survival was greater on average in the north (east = 9.8%, north = 10.8%, standard error of difference = 0.4%, $\chi^2_{(1)} = 6.05$, $P = 0.014$). Further analyses of adult components of fitness showed that lifetime reproductive success was higher (difference = $+0.182 \pm 0.047$ recruits, $t_{4243} = 3.92$, $P < 0.001$) and had increased significantly more over time in the north than in the east (difference = $+0.012 \pm 0.005$ recruits yr^{-1} of birth, $t_{4243} = 2.41$, $P = 0.016$). Thus, estimates of fitness components corroborate the suggestion that habitat quality differences drive differential dispersal and hence population differentiation. Other ecological mechanisms might have a role in the change in habitat quality, but these have not been quantified. Moreover, it is likely that many of these mechanisms would be correlated with the difference in density (for example, interspecific competition or food availability).

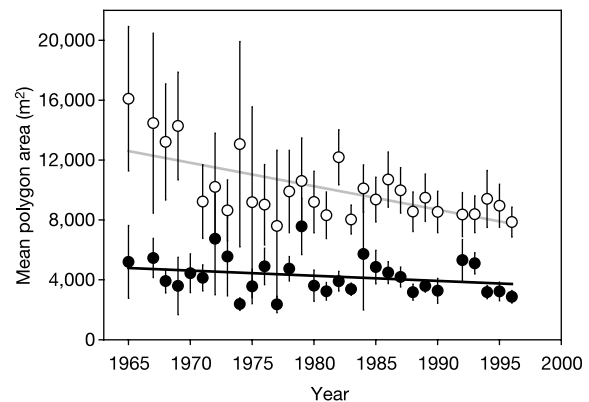


Figure 5 Mean polygon area (m^2) of breeding pairs of great tits in north (grey line, open circles) and east (black line, filled circles) areas over the study period (values are means $\pm$ 95% confidence intervals). This measure is inversely related to the breeding density experienced by breeding birds.

Here we have demonstrated significant evolutionary differentiation within a single population over small spatial and temporal scales. This differentiation results from the joint effects of differences in the expression of genetic variation and in dispersal, both potentially driven by variation in habitat quality. To put the patterns revealed here in context, the mean distance between breeding locations in the east and north populations is 3,720 m; the mean

natal dispersal distance of great tits in this population is 674 m for males (median 532 m) and 911 m for females (median 771 m)²⁰. The documented divergence here now amounts to 0.39 phenotypic standard deviations of fledging mass (change in haldanes: $h_{p(1.29)} = 0.02$), over a 36-year period (equivalent to 19 generations) and the average within-year quantitative genetic variation between areas (measured with Q_{st} on breeding value differences²¹) for the population between 1995 and 2000 is estimated to be 0.020 (95% confidence interval 0.003–0.037). These differences demonstrate that marked evolutionary differentiation is possible at small spatial and temporal scales within populations²², even in the absence of differences in selective regime²³. Our findings have important implications for future studies of local adaptation and speciation, and underline the importance of integrating knowledge about fine-scale environmental heterogeneity into such studies. □

Methods

Study species and data collection

We used data obtained from the long-term study of the great-tit population in Wytham Woods, Oxford, UK^{13,24}. All breeding attempts are monitored from the date of egg laying until all nestlings had fledged. At 15 days old (hatch day = 1), nestlings were weighed (to the nearest 0.1 g) and marked with individually numbered aluminium rings. At the same time, their parents were captured and their identity was checked, which allowed us to build pedigrees for the quantitative genetic analysis (see ref. 13 for further details). Here we used data from 1965 to 2000 because nest box locations were kept constant from 1964, and because we were able to correct for the effect of laying date, clutch size and egg weight for these years (see ref. 13). In all the analyses on fledging mass we therefore used the residual mass from a general linear model including laying date, clutch size and egg weight as variables. Second clutches and repeat clutches laid after failure of the first clutch were removed from the data set¹³. Thus, data were available for 4,856 breeding attempts involving 37,337 nestlings. To evaluate phenotypic trends of immigrants (from inside and outside) we used uncorrected adult mass (for birds caught in May); however, this was available for a more limited number of years (from 1974 to 2000 for birds from outside Wytham, $N = 1,185$; from 1978 to 2000 for birds from inside Wytham, $N = 217$). Linear regressions of annual means were weighted by the number of individuals in a given year in all cross-sectional analyses.

Quantitative genetic analyses

Variance components of fledging mass and individual breeding values were estimated through a mixed-model restricted-maximum likelihood (REML) estimation procedure with the software packages VCE4 (ref. 25) and PEST²⁶. We used the pedigree information to fit an individual 'animal model'¹⁸, in which year of birth was included as a random effect to account for temporal heterogeneity in environmental effects on the phenotype. The brood identity was also fitted as a random effect to account for common-environment and maternal effects specific to the individual brood. Best linear unbiased predictors of individual breeding values were then quantified from pedigree information by using REML estimates of variance components with the software package PEST. Change in estimated breeding values over different generations reflects changes in the genetic composition of the population. Because a lack of pedigree connectedness could introduce a form of gene–environment covariance, we calculated the proportion of individuals reproducing in either east or north (the areas with most contrasting changes in phenotype and genotype, and most separated in space) that had parents reproducing in the opposite habitat. We found that 9.2% of offspring reproduced in the habitat opposite to that of their parents (that is changed from north to east or vice versa within a generation). We also found that 72% of birds reproduced in the same habitat as their parents, but because each generation comprises two parents and given that we have a mean of six generations (excluding nestling generation) per individual in our pedigree, the probability that a lineage of the mean length is composed entirely of birds born in a given area is $(0.72^{26-1}) = 0.72^{63} = 1 \times 10^{-9}$.

Selection analyses

Selection intensity (i)¹ was estimated with linear regressions, for relative survival (within each area separately) on both standardized (zero mean, unit variance, within each area) phenotypic and breeding values of the body mass. Statistical significance of the standardized selection differentials was estimated with logistic regression²⁷. Here we used a logistic regression model with fixed (phenotype or genotypic values) and random effects that allowed us to take into account the non-independence of fledglings within nest boxes by including brood identity as a random effect in the model (a generalized linear mixed model, GLMM, with binomial error; implemented in Genstat version 7.1 (ref. 28)). The significance of body mass (or breeding value) as a predictor of survival ($0 =$ died, $1 =$ survived) was assessed from the Wald statistic, which is distributed as a χ^2_1 . Differences in chick survival between areas were tested by using a GLMM (with binomial error distribution) including area (east or north) and year (continuous) and their interactions as fixed effects and brood identity as a random effect. The association between area and adult lifetime reproductive success (LRS) (for adults born between 1965 and 1998; $N = 4,243$) was also tested by using a GLMM (with Poisson error distribution)

including adult status (immigrant or resident), area (east or north) and year of birth (continuous) and their interactions as fixed effects, and including year of birth as a random effect to account for cohort-specific variation (an analysis of adult LRS that also included adult mass as a covariate ($N = 2,607$) gave similar results).

Mapping

Average temporal changes in genotypic (breeding values) component of the fledging mass within each nest box (for which data were available for at least two years) were mapped (Fig. 2b) using inverse distance weighting interpolation. This method uses a moving average to interpolate pixel values and estimate local trends between spatially discontinuous and highly variable data. Each pixel value is calculated by averaging the weighted sums of all data points within a user-defined search area, such that points farther away influence the pixel value less than those that are close (decay exponent = 2). In this case, the radii of the search area and the display area were four times the average point density (413 m).

Density and definition of mean polygon area

To estimate the local breeding density experienced by each pair of great tits, Theissen polygons²⁹ were formed around each occupied nest box and their areas were calculated. Nest boxes were considered occupied if a confirmed breeding attempt was made by a pair of either great tits or blue tits (*Parus caeruleus*). We used data on both species (available from 1965 to 1996) because they compete for the same nest boxes and for food^{17,30}. In this case, Theissen polygons contain all points closer to a given nest box than to any other nest box, such that polygon boundaries are equidistant between occupied nest boxes. The area of each polygon is inversely related to the interspecific breeding density experienced by the breeding pair.

Received 2 July; accepted 21 September 2004; doi:10.1038/nature03051.

1. Endler, J. A. *Natural Selection in the Wild* (Princeton Univ. Press, Princeton, New Jersey, 1986).
2. Endler, J. A. *Geographic Variation, Speciation, and Clines* (Princeton Univ. Press, Princeton, New Jersey, 1977).
3. Slatkin, M. Gene flow and the geographic structure of natural populations. *Science* **236**, 787–792 (1987).
4. Felsenstein, J. The theoretical population genetics of variable selection and migration. *Annu. Rev. Genet.* **10**, 253–280 (1976).
5. García-Ramos, G. & Kirkpatrick, M. Genetic models of adaptation and gene flow in peripheral populations. *Evolution* **51**, 21–28 (1997).
6. Smith, T. B., Wayne, R. K., Gorman, D. J. & Bruford, M. W. A role for ecotones in generating rainforest biodiversity. *Science* **276**, 1855–1857 (1997).
7. Hendry, A. P., Day, T. & Taylor, E. B. Population mixing and the adaptive divergence of quantitative traits in discrete populations: a theoretical framework for empirical tests. *Evolution* **55**, 459–466 (2001).
8. Doebeli, M. & Dieckmann, U. Speciation along environmental gradients. *Nature* **421**, 259–264 (2003).
9. Barton, N. H. in *Dispersal* (eds Clobert, J., Danchin, E., Dhondt, A. A. & Nichols, J. D.) 329–340 (Oxford Univ. Press, 2001).
10. Verhulst, S., Perrins, C. M. & Riddington, R. Natal dispersal of great tits in a patchy environment. *Ecology* **78**, 864–872 (1997).
11. Braillet, C. et al. Two blue tit *Parus caeruleus* populations from Corsica differ in social dominance. *J. Avian Biol.* **33**, 446–450 (2002).
12. Naef-Daenzer, B., Widmer, F. & Nuber, M. Differential post-fledging survival of great and coal tits in relation to their condition and fledging date. *J. Anim. Ecol.* **70**, 730–738 (2001).
13. Garant, D., Kruuk, L. E. B., McCleery, R. H. & Sheldon, B. C. Evolution in a changing environment: a case study with great tit fledging mass. *Am. Nat.* **164**, E115–E129 (2004).
14. Merilä, J., Kruuk, L. E. B. & Sheldon, B. C. Cryptic evolution in a wild bird population. *Nature* **412**, 76–79 (2001).
15. Jensen, H. et al. Sexual variation in heritability and genetic correlations of morphological traits in house sparrow (*Passer domesticus*). *J. Evol. Biol.* **16**, 1296–1307 (2003).
16. Harvey, P. H., Greenwood, P. J. & Perrins, C. M. Breeding area fidelity of Great tits (*Parus major*). *J. Anim. Ecol.* **48**, 305–313 (1979).
17. Minot, E. O. & Perrins, C. M. Interspecific interference competition—nest sites for blue and great tits. *J. Anim. Ecol.* **55**, 331–350 (1986).
18. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer, Sunderland, Massachusetts, 1998).
19. Hoffman, A. A. & Merilä, J. Heritable variation and evolution under favourable and unfavourable conditions. *Trends Ecol. Evol.* **14**, 96–101 (1999).
20. van der Jeugd, H. P. & McCleery, R. Effects of spatial autocorrelation, natal philopatry and phenotypic plasticity on the heritability of laying date. *J. Evol. Biol.* **15**, 380–387 (2002).
21. Spitze, K. Population structure in *Daphnia obtusa*: Quantitative genetic and allozymic variation. *Genetics* **135**, 367–374 (1993).
22. Blondel, J., Dias, P. C., Perret, P., Maistre, M. & Lambrechts, M. M. Selection-based biodiversity at a small spatial scale in a low-dispersing insular bird. *Science* **285**, 1399–1402 (1999).
23. Postma, E. & van Noordwijk, A. J. Gene flow maintains a large genetic difference in clutch size at a small spatial scale. *Nature* doi:10.1038/nature03083 (this issue).
24. Perrins, C. M. Population fluctuations and clutch size in the great tit, *Parus major* L. *J. Anim. Ecol.* **34**, 601–647 (1965).
25. Neumaier, A. & Groeneveld, E. Restricted maximum likelihood estimation of covariances in sparse linear models. *Genet. Sel. Evol.* **30**, 3–26 (1998).
26. Groeneveld, E., Kovac, M., Wang, T. L. & Fernando, R. L. Computing algorithms in a general purpose BLUP package for multivariate prediction and estimation. *Arch. Anim. Breed.* **35**, 399–412 (1992).
27. Fairbairn, D. J. & Preziosi, R. F. Sexual selection and the evolution of sexual size dimorphism in the water strider, *Aquarius remigis*. *Evolution* **50**, 1549–1559 (1996).
28. VSN International, Genstat version 7.1 (VSN International, Oxford, 2003).
29. Rhynsbarger, D. Analytic delineation of Theissen polygons. *Geogr. Anal.* **5**, 133–144 (1973).

30. Minot, E. O. Effects of interspecific competition for food in breeding blue and great tits. *J. Anim. Ecol.* **50**, 375–385 (1981).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to A. Charmantier, A. G. Gosler, J. L. Quinn and C. M. Perrins for comments on the manuscript and to the many people who collected data during the long-term tit study in Wytham. D.G. was financially supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Postdoctoral Research Fellowship, and by a Biotechnology and Biological Sciences Research Council (BBSRC) grant to B.C.S. and L.E.B.K., who are both Royal Society University Research Fellows; T.A.W. was funded by a studentship from the BBSRC.

Authors' contributions D.G. conducted analyses and discovered the original pattern, and drafted the manuscript together with B.C.S., who also provided overall guidance, and L.E.B.K., who also advised over quantitative genetic analyses. T.A.W. conducted spatial analyses. R.H.McC. maintained the long-term database.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.G. (dany.garant@zoology.oxford.ac.uk).

Gene flow maintains a large genetic difference in clutch size at a small spatial scale

Erik Postma & Arie J. van Noordwijk

Netherlands Institute of Ecology (NIOO-KNAW), Centre for Terrestrial Ecology, PO Box 40, 6666 ZG Heteren, The Netherlands

Understanding the capacity of natural populations to adapt to their local environment is a central topic in evolutionary biology. Phenotypic differences between populations may have a genetic basis, but showing that they reflect different adaptive optima requires the quantification of both gene flow and selection^{1–3}. Good empirical data are rare⁴. Using data on a spatially structured island population of great tits (*Parus major*), we show here that a persistent difference in mean clutch size between two subpopulations only a few kilometres apart has a major genetic component. We also show that immigrants from outside the island carry genes for large clutches. But gene flow into one subpopulation is low, as a result of a low immigration rate together with strong selection against immigrant genes. This has allowed for adaptation to the island environment and the maintenance of small clutches. In the other area, however, higher gene flow prevents local adaptation and maintains larger clutches. We show that the observed small-scale genetic difference in clutch size is not due to divergent selection on the island, but to different levels of gene flow from outside the island. Our findings illustrate the large effect of immigration on the evolution of local adaptations and on genetic population structure.

Gene flow plays a crucial role in the evolution of natural populations^{2,3}. Although its role may be beneficial (by counteracting the negative effects of genetic drift and inbreeding on genetic variation^{5,6}), its role in constraining evolution by homogenizing the gene pool gains most attention^{1,7}. In spite of a wealth of theoretical studies showing that gene flow counteracts genetic differentiation, and thus the evolution of local adaptations, good empirical evidence is scarce, particularly for quantitative traits in spatially structured populations⁴. Those empirical studies that addressed the impact of gene flow on the evolution of local adaptations have focused on only some aspects, and investigated,

for example, the performance of populations inhabiting two different environments^{8,9}, or the effect of immigration into a single population^{10–12}. Furthermore, in the majority of studies assumptions had to be made regarding the genetic component in the phenotypic variation observed, or levels of selection and gene flow were not quantified (but see ref. 13).

A fascinating case of phenotypic variation at a small spatial scale can be found on the island of Vlieland in the Netherlands (Fig. 1a), where over the period 1975–95 females that bred in the western part of the island laid 1.15 ± 0.14 eggs more than females that bred in the eastern part (paired *t*-test: $t_{20} = 8.55$, $P < 0.001$), which is 0.91 s.d. of the average within-area and within-year distribution (Fig. 1b). Here we analyse what proportion of this persistent difference can be attributed to genetic variation, and then investigate the roles of immigration and several components of selection in maintaining this difference.

The classic approach to test for local adaptations, and genetic differences in general, is a common garden or other transplant experiment. Although such experiments are relatively easy to perform in a laboratory setting^{14,15}, or in sessile organisms in the wild¹⁶, they are not feasible for the majority of wild animals (but see refs 13, 17). However, about 10% of the females born on Vlieland breeds in the other area (also, see below), which allows for a separation of genetic and environmental effects. Additionally, the properties of this long-term study population allow for an accurate estimation of both immigration and selection (see Methods).

Whereas the difference in clutch size between the western and eastern part of the island ('East' and 'West', respectively) can partly be attributed to the area of breeding, and thus to phenotypic plasticity, approximately half of the difference is accounted for by where a female is born, which suggests a genetic component to the difference in clutch size between the West and the East (Fig. 2a). Although females that have dispersed to the other side of the island may either represent a non-random subset of birds or breed, for

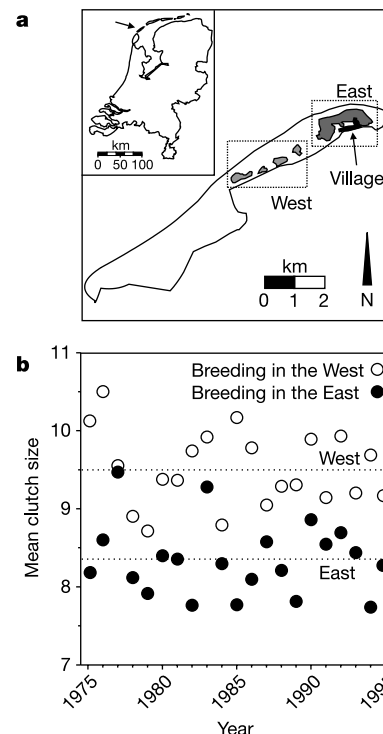


Figure 1 Clutch sizes on the island of Vlieland. **a**, Map of Vlieland, with the location of the woodlands that provide suitable nesting habitat in grey, and their grouping into West and East. **b**, Yearly mean clutch size in the West and the East from 1975 to 1995. Dotted lines give the overall mean.

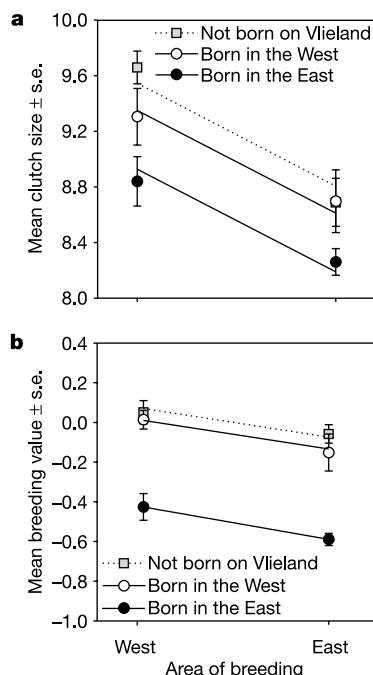


Figure 2 Spatial variation in clutch size. **a**, Mean clutch sizes of females in relation to origin and area of breeding (born in the West or the East: origin: $F_{1,67} = 6.43$, $P = 0.014$; area of breeding: $F_{1,67} = 11.0$, $P = 0.002$, interaction: $F_{1,66} = 0.01$, $P = 0.94$). Immigrants do not differ significantly from females born in the West: ($t_{108} = 1.15$, $P = 0.25$), but they do from females born in the East (after Tukey–Kramer adjustment for multiple comparisons: $t_{108} = 3.86$, $P = 0.006$). **b**, Mean predicted breeding values for clutch size for males in the same categories (born in the West or the East: origin: $F_{1,56} = 57.9$, $P < 0.001$; area of breeding: $F_{1,56} = 8.11$, $P = 0.006$). Immigrants do not differ significantly from males born in the West: ($t_{82} = 1.03$, $P = 0.31$), but they do from males born in the East (after Tukey–Kramer adjustment for multiple comparisons: $t_{82} = 10.02$, $P < 0.001$).

example, in poorer territories, this should affect birds from the West and the East similarly, and result in crossing instead of parallel reaction norms. Moreover, we find a very similar difference in the mean predicted breeding values for clutch size of males (see Methods) born in the West and the East (Fig. 2b), which argues against an environmental effect of the area of birth lasting throughout an individual's life. On the whole, we have strong evidence for a genetic difference in clutch size of approximately 0.5 egg between birds born in the West and the East.

To understand how this large genetic difference could be maintained over a period of 21 yr at such a small spatial scale, and more specifically whether this difference reflects different adaptive optima, we first compared the viability and fecundity of birds born in the West and the East breeding in either of the two areas. This provides us with the most direct and general test for local adaptation to the environment in the West and the East. Although both males and females produce about twice as many recruits (see Methods) when they breed in the East than in the West (males: year: $F_{20,36} = 2.59$, $P = 0.006$; area of breeding: $F_{1,36} = 20.86$, $P < 0.001$; females: year: $F_{20,47} = 4.18$, $P < 0.001$; area of breeding: $F_{1,47} = 11.9$, $P = 0.001$), there is no significant difference between birds born in the West and the East (males: $F_{1,35} = 0.005$, $P = 0.95$ (Fig. 3a); females: $F_{1,46} = 2.79$, $P = 0.10$). Local survival (see Methods) of females born in the East is higher than that of females born in the West, irrespective of where they breed (Fig. 3b). The difference in local survival between males born in the West and the East is not significant (year: $F_{1,57} = 1.04$, $P = 0.44$; area of breeding: $F_{1,57} = 8.18$, $P = 0.021$; area of birth: $F_{1,57} = 1.38$, $P = 0.18$), and is significantly different from that in females (model including both

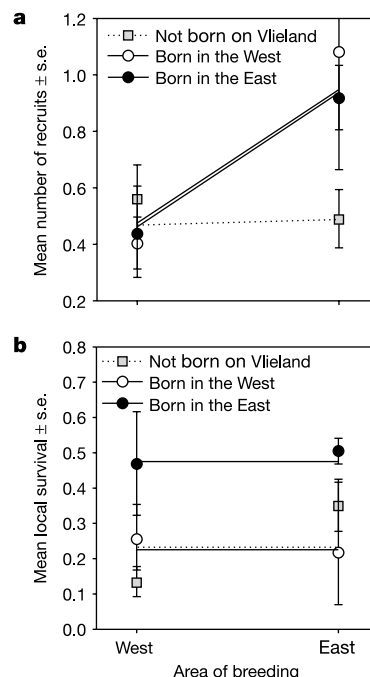


Figure 3 Variation in fitness in relation to origin and area of breeding. **a**, Mean number of recruits produced in one year for local and immigrant males, and breeding in the West or the East (year: $F_{20,54} = 3.25$, $P < 0.001$; local or immigrant $\times$ area of breeding: $F_{1,54} = 11.3$, $P = 0.001$; Local versus immigrant in the West: $t_{54} = 1.37$, $P = 0.18$, and in the East: $t_{54} = 3.32$, $P = 0.002$). **b**, Mean local survival of females (born in the West or the East: year: $F_{20,48} = 1.88$, $P = 0.037$; origin: $F_{1,48} = 6.61$, $P = 0.013$; area of breeding: $F_{1,47} = 0.04$, $P = 0.85$). Immigrants do not differ significantly from females born in the West: ($t_{89} = 0.078$, $P = 0.94$), but they do from females born in the East (after Tukey–Kramer adjustment for multiple comparisons: $t_{89} = 2.86$, $P = 0.015$).

sexes: sex $\times$ area of birth: $F_{1,103} = 8.22$, $P = 0.005$). So, local survival of females born in the East is twice as high as that of females born in the West, both in the West and the East. This indicates that females born in the East are better adapted to the environment on Vlieland in general than females born in the West, and argues against local adaptation of birds to their area of birth.

To investigate the role of immigration from outside Vlieland, we compared their clutch size and fitness to birds born in the West and the East, and quantified the number of immigrants into both parts of the island. The inclusion of immigrants into the analyses above shows that immigrant females lay significantly larger clutches than females born in the East (Fig. 2a), and that immigrant males have significantly higher predicted breeding values for clutch size than males born in the East (Fig. 2b). Furthermore, immigrant females have a significantly lower local survival than females born in the East (Fig. 3b). In neither of these cases, however, can they be distinguished from birds born in the West. Finally, when immigrant males breed in the East they produce significantly fewer recruits than males born in both the East ($t_{68} = 3.03$, $P = 0.0035$) and the West ($t_{68} = 2.15$, $P = 0.035$) (Fig. 3a). In females this difference in recruitment between immigrants and locals in the East is absent (born on Vlieland or not $\times$ area of breeding: $F_{1,60} = 0.05$, $P = 0.82$), and significantly different from the difference between local and immigrant males (model including both sexes: sex $\times$ born on Vlieland $\times$ area of breeding: $F_{1,134} = 4.76$, $P = 0.031$).

The fact that birds born in the West are genetically very similar to immigrants, at least with respect to clutch size, can be understood from the fact that the proportion of first-year breeders that is not born on Vlieland is on average 3.3 times higher in the West (43%) than in the East (13%) (Fig. 4a). This, together with the relatively

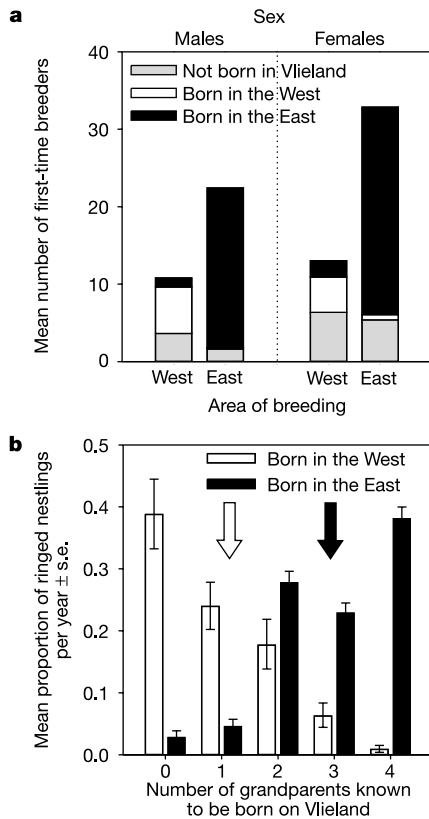


Figure 4 Immigration into the East and the West. **a**, The average composition of the breeding population for both the West (mean proportion of immigrants among first-time breeders = 43%, 95% CI: 35–51%) and the East (mean = 13%, 95% CI: 10–16%). **b**, The proportion of all ringed nestlings per year against the number of local grandparents for both the West (mean across all young ringed as nestlings = 1.07 ± 0.098 ; open arrow) and the East (mean = 2.85 ± 0.049 ; filled arrow).

higher recruitment of immigrants in the West, results in a much higher level of gene flow into this part of the island, and as a consequence birds that are born in the West are more closely related to immigrants than birds born in the East (Fig. 4b).

On the basis of these findings, we propose that the following mechanism underlies the genetic difference between birds born in the West and the East: immigrants carry the genes for relatively large clutches. Both in the East and the West there is selection against immigrant genes in females, and in the East there is additional selection against immigrant males. In the Eastern sub-population this, together with the lower immigration rate, is maintaining local adaptation to the environment on Vlieland, as well as smaller clutches. The higher level of gene flow in the West results in larger clutches, and prevents local adaptation. This mechanism is supported by two findings, which are directly predicted from Figs 2b and 3b, respectively: within the males born on Vlieland, those males that are less related to immigrants have genes for smaller clutches (Fig. 5a). Furthermore, of the females born on Vlieland, those females that are less related to immigrants have a higher local survival (Fig. 5b).

We have shown that great tits not born on Vlieland have genes for larger clutches, and have a lower survival on Vlieland. However, within the East there is no evidence for viability selection against large clutches when we correct for a female's origin (year: $\chi^2_{(20)} = 61.8$, $P < 0.001$; area of birth: $\chi^2_{(2)} = 6.09$, $P = 0.048$; standardized clutch size: $\chi^2_{(1)} < 0.005$, $P = 0.96$, $|\beta'| < 0.001$) (see Methods), which argues against a causal relationship between the two. The lower survival of immigrant genotypes is thus not directly related to clutch size. Furthermore, across the period

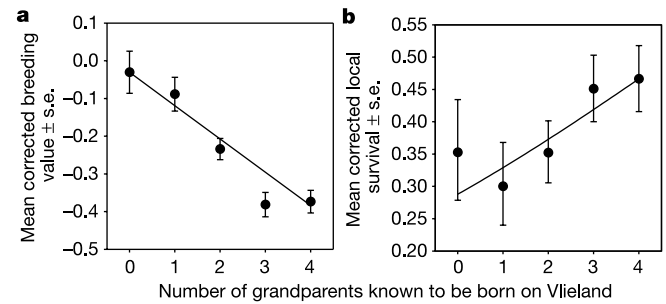


Figure 5 Clutch size and fitness in relation to an individual's relatedness to immigrants. **a**, The relationship between predicted breeding values for clutch size of males and the number of local grandparents ($b = -0.088 \pm 0.015$, $F_{1,535} = 36.5$, $P < 0.001$), after correction for an effect of origin ($F_{1,535} = 23.8$, $P < 0.001$) and the area of breeding ($F_{1,535} = 5.42$, $P < 0.001$). **b**, The relationship for females between local survival and the number of local grandparents ($\chi^2_{(1)} = 4.98$, $P = 0.027$), corrected for an effect of year ($\chi^2_{(20)} = 53.69$, $P < 0.001$) and area of birth ($\chi^2_{(1)} = 8.10$, $P = 0.004$).

1975–95 there is no relationship between clutch size and the number of recruits produced (year: $\chi^2_{(20)} = 126.2$, $P < 0.001$; standardized clutch size: $\chi^2_{(1)} = 0.63$, $P = 0.43$, $|\beta'| < 0.001$). Across the years before 1975, however, there is a clear trend that females that laid small clutches produced more recruits (year: $\chi^2_{(15)} = 61.5$, $P < 0.001$; standardized clutch size: $\chi^2_{(1)} = 3.46$, $P = 0.063$, $\beta' = -0.14$). This relatively strong selection¹⁸ could explain how the small clutches of birds born in the East have evolved in the first place, but also the potential for founder effects should not be ruled out.

Great tits on Vlieland maintain their small clutches against 13%, but not against 43%, immigration. Although phenotypic variation across environments is often assumed to be adaptive¹⁹, we show that a scenario involving different levels of gene flow can maintain genetic differences in a life-history trait for a highly mobile animal at a very small spatial scale (see also ref. 20). Our results illustrate the major consequences of immigration for genetic population structure and fitness. Even though Vlieland is relatively isolated when compared to many mainland populations²¹, within-island movements have a negligible effect. The properties of the Vlieland population have allowed us to disentangle the role of genetic and environmental variation, and of gene flow and selection. There is, however, no reason to believe that similar processes would not take place in many more populations where they would remain undetected. □

Methods

Data set

The data were collected from the nest-box breeding population of great tits on the island of Vlieland (53.17° N, 5.03° E), The Netherlands. The first nest boxes were put up in 1955, and the population has been monitored continuously since then. On the basis of exchange rates among the five separate woodlands, the population can be divided into an eastern and a western sub-population^{22,23}. Both subpopulations are at least 1.3 km, and on average 5.1 km, apart (Fig. 1a). Unless specified otherwise, analyses are performed on first clutches of first year breeders from 1975 until and including 1995. Clutch size is assumed to be a trait of the laying female (see Supplementary Information). Manipulated clutches were included in the analyses (see Supplementary Information). There was no temporal trend in mean clutch size over this period, either in the West (linear regression: $b = -0.013 \pm 0.018$, $t_{19} = -0.72$, $P = 0.48$), or in the East ($b = -0.010 \pm 0.017$, $t_{19} = -0.56$, $P = 0.58$) (Fig. 1b).

Nest boxes are provided in excess in all suitable nesting habitat. Females breeding in either the East or the West that had not been ringed as a nestling anywhere on Vlieland are considered to be immigrants, although a small proportion of those may have been born on Vlieland²¹. This will, however, make the majority of our tests more conservative. The degree of relatedness of birds born on Vlieland to immigrants is expressed in how many of their four grandparents had been ringed as nestlings on Vlieland. Immigrants are assumed to have parents who were not born on Vlieland. Birds for which the origin of any of their parents or grandparents could not be inferred were excluded. The mean number of grandparents born on Vlieland of birds born in the East or the West did not change significantly over time (linear regression: $b = 0.0047 \pm 0.010$, $t_{19} = 0.46$, $P = 0.65$).

Statistical analyses

To account for the non-independence of individual observations within a year, and the large differences in sample size among groups, analyses were, where possible, performed on yearly means for the relevant sub-group. When appropriate, yearly means were square-root or arc-sine transformed before analysis²⁴. All presented means and parameter estimates are back-transformed.

All statistical analyses were performed using the SAS statistical package using the GLM and GENMOD procedures²⁵. Non-significant interactions were removed first, starting with the least significant, followed by non-significant single terms, again starting with the least significant. When interactions were significant, the main effects were kept in the model, regardless of their significance. All tests are two-tailed.

Selection analyses

Local survival was used as a measure of viability, which is defined as the probability that a bird is observed breeding again on Vlieland the next year. Local recruitment was used as a measure of fecundity, and is defined as the number of offspring produced in all clutches in a year observed to be breeding on Vlieland in subsequent years. We can therefore not distinguish between mortality and emigration of both fledglings and adults. Both processes do however have a similar effect on the population level. Only clutches from which at least one chick fledged were included in analyses of recruitment.

To quantify selection acting on clutch size, we calculated standardized selection gradients (β') by regressing relative fitness (fitness divided by the mean fitness in that year) on standardized clutch size (clutch size relative to the mean in that year, divided by the standard deviation)²⁶. Significance of selection gradients was determined from a generalized linear model with binomial and Poisson errors, using standardized clutch sizes. If significant, origin of birth was included in the model. Selection gradients were calculated for both the period from 1955 until 1975, and for 1975 until 1995. As a result of the small sample sizes in the West, especially in the first period, analyses were limited to females breeding in the East.

Animal model analysis

Genetic parameters were estimated using an animal model, which uses all available information on both ancestors and descendants to separate an individual's phenotype into an additive genetic component (or breeding value) and other random and fixed effects^{27,28}. The amounts of variance accounted for by the random effects (additive genetic variance V_A , permanent environmental variance V_{PE} , and residual variance V_R) were estimated using a Restricted Maximum Likelihood (REML) technique implemented in the software package VCE4²⁹, and were equal to 0.58, 0.42 and 0.94, respectively. The narrow sense heritability $\pm$ s.e. of clutch size on Vlieland across the period 1965–2003 (defined as V_A/V_P) was 0.30 ± 0.028 . Best Linear Unbiased Predictions (BLUPs) of breeding values were obtained for all individuals in the pedigree using the software package PEST³⁰. See Supplementary Information for more details on pedigree reconstruction and which fixed and random effects were included in the animal model.

Received 16 September; accepted 8 October 2004; doi:10.1038/nature03083.

1. Slatkin, M. Gene flow in natural populations. *Annu. Rev. Ecol. Syst.* **16**, 393–430 (1985).
2. Bohonak, A. J. Dispersal, gene flow, and population structure. *Q. Rev. Biol.* **74**, 21–45 (1999).
3. Lenormand, T. Gene flow and the limits to natural selection. *Trends Ecol. Evol.* **17**, 183–189 (2002).
4. Hendry, A. P., Day, T. & Taylor, E. B. Population mixing and the adaptive divergence of quantitative traits in discrete populations: A theoretical framework for empirical tests. *Evolution* **55**, 459–466 (2001).
5. Ebert, D. *et al.* A selective advantage to immigrant genes in a *Daphnia* metapopulation. *Science* **295**, 485–488 (2002).
6. Keller, L. F. *et al.* Immigration and the ephemerality of a natural population bottleneck: evidence from molecular markers. *Proc. R. Soc. Lond. B* **268**, 1387–1394 (2001).
7. Ehrlich, P. R. & Raven, P. H. Differentiation of populations. *Science* **165**, 1228–1232 (1969).
8. King, R. B. & Lawson, R. Color-pattern variation in Lake Erie water snakes: The role of gene flow. *Evolution* **49**, 885–896 (1995).
9. Dhondt, A. A., Adriaenssen, F., Matthysen, E. & Kempenaers, B. Nonadaptive clutch sizes in tits. *Nature* **348**, 723–725 (1990).
10. Clobert, J., Perrins, C. M., McCleery, R. H. & Gosler, A. G. Survival rate in the great tit *Parus major* in relation to sex, age, and immigration status. *J. Anim. Ecol.* **57**, 287–306 (1988).
11. McCleery, R. H. & Clobert, J. in *Population Biology of Passerine Birds* (eds Blondel, J., Gosler, A. G. & Clobert, J.) 423–440 (Springer, Berlin, 1990).
12. Marr, A. B., Keller, L. F. & Arcepe, P. Heterosis and outbreeding depression in descendants of natural immigrants to an inbred population of song sparrows (*Melospiza melodia*). *Evolution* **56**, 131–142 (2002).
13. Hendry, A. P., Taylor, E. B. & McPhail, J. D. Adaptive divergence and the balance between selection and gene flow: Lake and stream stickleback in the Misty system. *Evolution* **56**, 1199–1216 (2002).
14. Stearns, S. C. & Sage, R. D. Maladaptation in a marginal population of the mosquito fish, *Gambusia affinis*. *Evolution* **34**, 67–75 (1980).
15. Storfer, A. & Sih, A. Gene flow and ineffective antipredator behavior in a stream-breeding salamander. *Evolution* **52**, 558–565 (1998).
16. Stanton, M. L. & Galen, C. Life on the edge: Adaptation versus environmentally mediated gene flow in the snow butterfly, *Ranunculus adoneus*. *Am. Nat.* **150**, 143–178 (1997).
17. Lambrechts, M. M. & Dias, P. C. Differences in the onset of laying between island and mainland Mediterranean blue tits *Parus caeruleus*—Phenotypic plasticity or genetic differences. *Ibis* **135**, 451–455 (1993).
18. Kingsolver, J. G. *et al.* The strength of phenotypic selection in natural populations. *Am. Nat.* **157**, 245–261 (2001).
19. Krebs, J. R. & Davies, N. B. *Behavioural Ecology: An Evolutionary Approach* (Blackwell, Oxford, 1991).
20. Garant, D., Kruuk, L. E. B., Wilkin, T. A., McCleery, R. H. & Sheldon, B. C. Evolution driven by differential dispersal within a wild bird population. *Nature* doi:10.1038/nature03051 (this issue).

21. Verhulst, S. & Van Eck, H. M. Gene flow and immigration rate in an island population of great tits. *J. Evol. Biol.* **9**, 771–782 (1996).
22. Van Noordwijk, A. J., Van Balen, J. H. & Scharloo, W. Genetic and environmental variation in clutch size of the great tit (*Parus major*). *Neth. J. Zool.* **31**, 342–372 (1981).
23. Van Tienderen, P. H. & Van Noordwijk, A. J. Dispersal, kinship and inbreeding in an island population of the great tit. *J. Evol. Biol.* **1**, 117–137 (1988).
24. Quinn, G. P. & Keough, M. J. *Experimental Design and Data Analysis for Biologists* (Cambridge Univ. Press, Cambridge, 2002).
25. SAS Institute Inc., *SAS/STAT User's Guide* (SAS Institute Inc., North Carolina, 1996).
26. Arnold, S. J. & Wade, M. J. On the measurement of natural and sexual selection: Theory. *Evolution* **38**, 709–719 (1984).
27. Kruuk, L. E. B. Estimating genetic parameters in natural populations using the 'animal model'. *Phil. Trans. R. Soc. Lond. B* **359**, 873–890 (2004).
28. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer Associates, Massachusetts, 1998).
29. Neumaier, A. & Groeneveld, E. Restricted maximum likelihood estimation of covariances in sparse linear models. *Genet. Sel. Evol.* **30**, 3–26 (1998).
30. Groeneveld, E., Kovac, M., Wang, T. L. & Fernando, R. L. Computing algorithms in a general-purpose BLUP package for multivariate prediction and estimation. *Arch. Anim. Breed.* **35**, 399–412 (1992).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This study would have been impossible without all those who collected data on Vlieland, and H. van Eck in particular. J. Visser maintained the database. L. Keller, K. Lessells, K. van Oers and M. Visser provided comments on earlier versions of the manuscript. E.P. is supported by ALW-NWO and a Marie-Curie fellowship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.P. (e.postma@nioo.knaw.nl).

A mechanism for impaired fear recognition after amygdala damage

Ralph Adolphs^{1,2}, Frederic Gosselin³, Tony W. Buchanan¹, Daniel Tranel¹, Philippe Schyns⁴ & Antonio R. Damasio¹

¹Department of Neurology, University of Iowa, Iowa City, Iowa 52242, USA

²Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, California 91125, USA

³Département de Psychologie, Université de Montréal, Montréal H3C 3J7, Canada

⁴Department of Psychology, University of Glasgow, Glasgow G12 8QB, UK

Ten years ago, we reported that SM, a patient with rare bilateral amygdala damage, showed an intriguing impairment in her ability to recognize fear from facial expressions¹. Since then, the importance of the amygdala in processing information about facial emotions has been borne out by a number of lesion^{2–4} and functional imaging studies^{5,6}. Yet the mechanism by which amygdala damage compromises fear recognition has not been identified. Returning to patient SM, we now show that her impairment stems from an inability to make normal use of information from the eye region of faces when judging emotions, a defect we trace to a lack of spontaneous fixations on the eyes during free viewing of faces. Although SM fails to look normally at the eye region in all facial expressions, her selective impairment in recognizing fear is explained by the fact that the eyes are the most important feature for identifying this emotion. Notably, SM's recognition of fearful faces became entirely normal when she was instructed explicitly to look at the eyes. This finding provides a mechanism to explain the amygdala's role in fear recognition, and points to new approaches for the possible rehabilitation of patients with defective emotion perception.

Patient SM is a 38-yr-old woman whose brain lesion encompasses all nuclei of the amygdala bilaterally, as well as a small portion of the

adjacent entorhinal cortex, yet spares all other subcortical and cortical structures, leaving her with essentially normal basic perception, memory, language and reasoning insofar as these do not involve the processing of emotional material⁷. However, her processing of emotionally and socially meaningful information is impaired, as it is in nonhuman animals with amygdala damage. For example, she does not show normal conditioned fear responses⁸, and her social behaviour is indiscriminately trusting and friendly⁹. Over more than a decade of testing, she has consistently shown a severe and selective impairment in the ability to recognize fear from facial expressions^{1,7}, although she is able to recognize fear from complex visual scenes and tone of voice. So far, she remains the human subject with the most selective amygdala damage and with the most selective impairment in fear recognition from faces; however, no mechanism has yet been provided to link these two conditions.

We began by exploring SM's ability to make use of visual information from specific regions of the face. SM and normal control subjects were each shown approximately 3,000 trials of sparsely revealed faces varying in gender and emotional expression (fear or happiness)^{10,11}. In each trial, random locations on one of the face images were made visible with gaussian 'bubbles' in five one-octave bands of spatial frequencies (see Supplementary Fig. 1), and viewers were asked in a two-alternative discrimination task to judge whether the revealed features expressed fear or happiness. We chose to contrast these two expressions because SM differs most in her ability to recognize them (entirely normal recognition of happiness, severely impaired recognition of fear)^{1,7}, and because they differ most in terms of the facial features used for their identification¹². For each subject, recognition performance was kept constant at 75% for each emotion by interactively adjusting the number of bubbles during the task. This corresponded to an average of 16.5 bubbles (s.d. = 3.1, range = 13–23.4) per image for the normal controls, whereas SM required 30.8 bubbles per image. The number of bubbles required to identify correctly a face as fearful or happy was equivalent; the difference in number of bubbles (fearful faces minus happy faces) was -0.03 bubbles for control subjects and $+0.05$ bubbles for SM.

Is SM's requirement for more bubbles relative to control subjects due to a decrease in her use of visual information over all facial features, or can it be attributed to a failure in using information from specific facial features? We performed a linear regression using the location of the bubbles on the face images and the subject's discrimination accuracy on each trial to reveal the regions of the face used to discriminate between fear and happiness. Whereas normal subjects used information predominantly from the eyes in high spatial frequencies (from 5.59–22.38 cycles per degree), SM failed to make the same use of eye information (Fig. 1a, b). For the highest spatial frequency band information from the eyes, SM's mean Z-score was equal to 0.59 s.d. below her global mean (that is, her use of the eyes at high spatial frequency was worse than her mean use of all face regions across all spatial frequencies), whereas the Z-scores of control subjects ranged from 0.42 to 1.50 s.d. above the mean (average = $+0.79$). Whereas every normal subject made use of visual information from the eye region in the highest spatial frequency band ($P < 0.05$), SM did not.

Moreover, SM did not use information in the face other than the eyes more effectively than control subjects when discriminating fear; the difference image of the visual information used more by SM than by control subjects does not reveal any such features (Fig. 1b). Although SM failed to use information from the eyes in high spatial frequencies in gaussian bubble trials showing either fearful or happy faces, she did make normal use of the mouth region (Fig. 1c). This finding probably explains her intact ability to recognize happiness, and her equivalent performance at discriminating between fearful and happy faces in the described task—because we offered her only two options, her intact ability to use the smile to identify

happiness should result in successful identification of fear by exclusion.

SM's failure to use information about the eyes stood out as abnormal in comparison with every one of the ten normal control subjects we tested (Supplementary Figs 2 and 3). In order to establish further the specificity of SM's deficit, we performed the same two-alternative discrimination task in 13 subjects with unilateral amygdala damage and with normal fear recognition. All made normal use of information from the eye region of the faces (see Supplementary Fig. 4).

Although the large number of trials required precluded testing SM's ability to discriminate fear from all other basic emotions on this particular task, such data have been obtained in a separate study in normal individuals¹². When asked to discriminate between each of the six basic emotions (happiness, surprise, fear, anger, disgust and sadness) and neutral expressions in a seven-alternative discrimination task, normal subjects consistently and specifically make the most use of high spatial frequency information from the eyes for discriminating fear. It is interesting to note that discrimination of

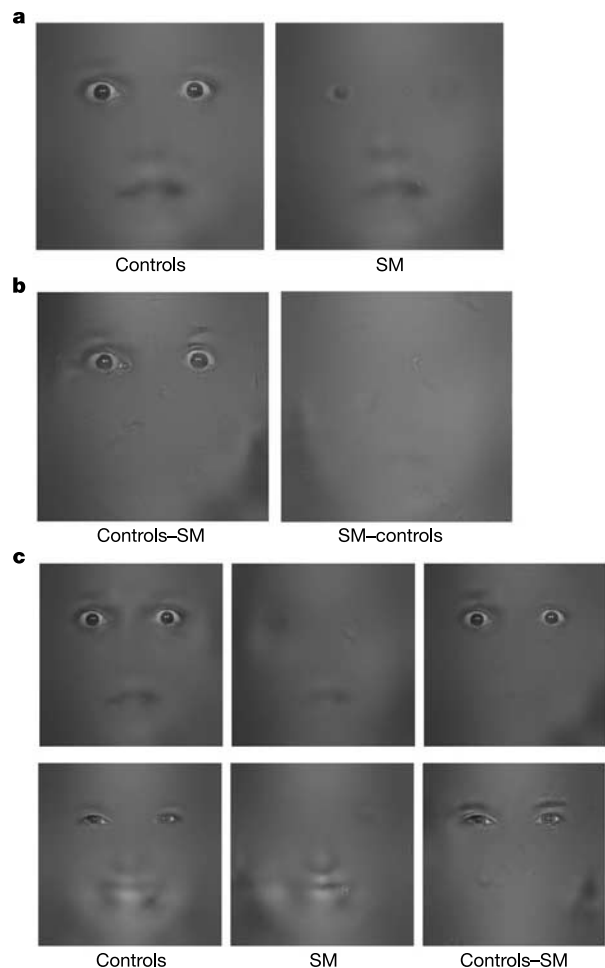


Figure 1 SM fails to make use of visual information from the eyes in faces. **a**, Information from faces used to discriminate fear from happiness in ten control subjects (left panel) or SM (right panel). **b**, Difference images showing the facial information used more by control subjects than by SM (left panel), or more by SM than by control subjects (right panel). Unlike control subjects, SM does not use high spatial frequency information about the eyes, nor does she use any information that the controls do not. **c**, Visual information used in those trials in which fearful faces were shown (top row) or happy faces were shown (bottom row). SM fails to make use of the eyes for either emotion, but is able to use information about the mouth normally.

two other emotions, sadness and anger, also makes substantial use of the eye region, and that recognition of these two emotions, in addition to fear, has been most consistently reported to be impaired after amygdala damage in other patients³. The highly selective impairment in fear recognition in SM's case is probably attributable to her ability to make compensatory use of information outside the eye region for those other emotions; however, this strategy is insufficient in the case of fear.

In a control task using identical stimuli and procedure to those described above, subjects were asked to discriminate the gender of the faces rather than their emotion. SM's performance was normal in all respects for this task: she required the same number of bubbles (average number required by control subjects = 46.5, s.d. = 9.5; number required by SM = 39.5) and she used exactly the same effective visual information (the difference image for control subjects minus SM was uniformly grey). Notably, both SM and controls

used high spatial frequency information from the eyes and mouth in the gender discrimination task (see Supplementary Fig. 5), indicating that SM is indeed capable of using such information, although she fails to do so spontaneously when judging emotion.

The discrimination task using the gaussian bubbles method provided an unbiased and homogeneous sampling of all regions of the face that might be important for fear recognition, but used rather artificial stimuli that might be processed differently than actual faces, and was restricted to comparisons between two emotions (fear and happiness). We thus conducted a further experiment to assess directly the importance of the eyes within facial images and broaden the scope of our conclusions. Subjects were shown whole facial images expressing the six basic emotions, as well as the same images with the eyes digitally erased, and we assessed their accuracy in recognizing the emotion in each image. Whereas control subjects were significantly less accurate at recog-

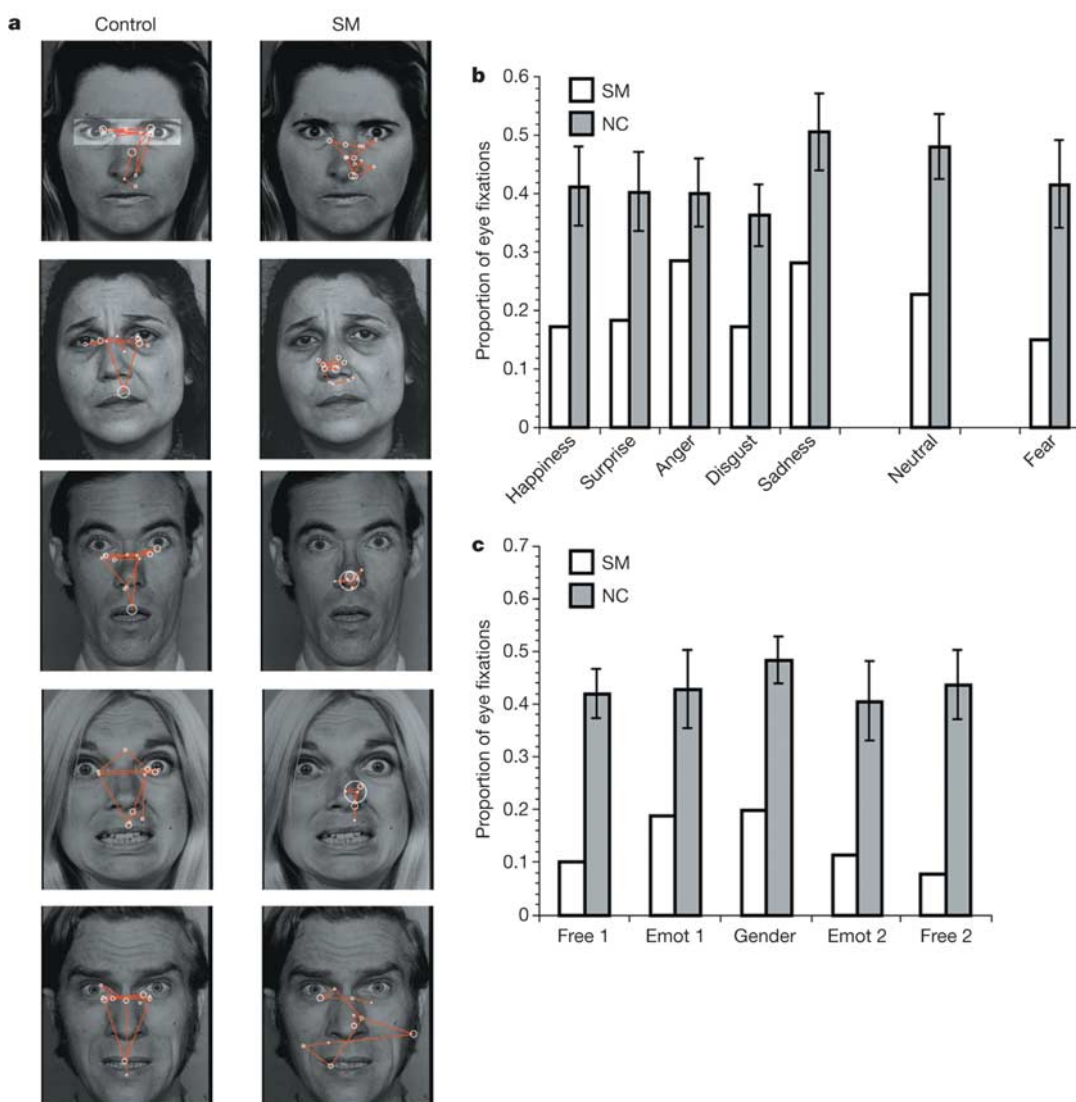


Figure 2 SM fails to fixate on the eyes when viewing facial expressions. **a**, Saccades (red lines) and fixations (white circles, where circle size corresponds to fixation duration) made by a typical normal control subject (left column) and SM (right column) when judging the emotion shown in sample expressions (from top to bottom) of anger, sadness and three fear faces. A lightly shaded box around the eyes is present in the top left image, showing the region (defined *a priori*) used to calculate the proportion of fixations shown in **b**, **b**, The proportion of fixations made by SM (white bars) and normal control subjects (NC, grey

bars, mean $\pm$ s.e.m.) on the eye region of face images when judging different emotions, calculated as the number of fixations to the eye region divided by the total number of fixations made on the face. **c**, The proportion of fixations made specifically to facial expressions of fear, under the five different viewing conditions detailed in the Methods, shown in their order of presentation from left to right (Free = passive viewing, Emot = emotion judging). SM's proportion of fixations on the eyes is abnormally low for all conditions.

nizing fear when the eyes had been erased ($P < 0.005$, paired t -test), SM showed no change in her performance accuracy (0.33 in both conditions). No control subject ever approached SM's performance in fear recognition for whole faces (lowest control performance of 0.67 accuracy) whereas three out of twelve control subjects were as impaired as or worse than SM when the eyes had been erased. Notably, this pattern extended to other emotions (see Supplementary Table 1), as the recognition accuracy of control subjects dropped when the eyes were erased, but SM's accuracy did not. These findings confirmed that SM fails to make normal use of information from the eye region of faces when judging facial emotions.

The findings thus far raised the possibility that SM's impairment might result from a failure to direct her gaze to the eyes in the first place. To test this idea, we monitored eye movements while subjects viewed prototypical facial expressions of all basic emotions^{13,14} under five conditions: passive viewing (done twice), emotion recognition (done twice) and gender recognition (done once). Normal control subjects reliably explored the face, fixating mostly on the eyes (Fig. 2a); this is a pattern observed in humans as young as 7 weeks old¹⁵ as well as in nonhuman primates¹⁶. SM showed a highly abnormal fixation pattern: she did not explore the face normally, and systematically failed to fixate on the eye region. This impairment was evident for fear as well as other emotions (Fig. 2b). SM's fixations on the eyes were fewer than those of any normal control subject, and were significantly fewer than the control group for all but one condition (the first emotion judgement task 'Emot 1' in Fig. 2c, $P < 0.2$; all other conditions, $P < 0.05$; two-tailed Z -tests).

A control task verified that SM's abnormal fixations do not arise from cueing particular locations during the experimental procedure. Specifically, the fixation cross that preceded each face stimulus in the above experiments was located in the centre of the screen, roughly coincident with the subsequent location of the nose

in each face. A further two blocks of trials presented the same faces, but preceded by a fixation cross coincident with either the left or right eye rather than the nose, and asked subjects to judge the emotion. SM's proportion of fixations to the eyes remained abnormally low (0.24 for both trial blocks versus 0.49 and 0.48 respectively for the control subjects), and her fear recognition remained impaired (0.33 and 0.17 correct for the two trial blocks versus 0.81 and 0.79 for the control subjects).

We interpreted the above findings to mean that SM is impaired in recognizing fear because she is unable to make use of diagnostic information from the eye region that is normally essential for recognizing fear, and that this inability is related to her lack of spontaneous fixation on the eye region of faces. This interpretation would predict that manipulating how she inspects faces might influence her ability to recognize emotion. Accordingly, we reassessed her emotion recognition while instructing her specifically to look at the eye region of faces. As instructed, SM looked at the eyes in the facial expressions presented (Fig. 3). Her impaired recognition of fear was completely reversed (that is, attained normal levels) with this simple instruction. We verified this result on two separate occasions, counterbalancing the order of the 'instruction' task and the previously described free viewing task (Fig. 3 and Table 1).

However, a single instruction to direct her gaze onto the eye region of facial images was insufficient to rehabilitate permanently SM's impaired fear recognition. When we subsequently showed her the face stimuli under unconstrained viewing conditions, she failed to fixate the eye region spontaneously and reverted to her previously impaired fear recognition. Thus the impairment could be rescued by instruction to fixate the eye region of faces, but the improvement lasted only as long as the instruction remained explicit. This finding opens the possibility for developing a strategy that could consistently direct her gaze to the eye region of faces, perhaps with additional instruction and training.

In over a decade of repeated testing, SM has not learned to recognize fear in faces⁷, and does not appear to have improved her defective social judgements⁹. This collection of impairments is consistent with an inability to search automatically for environmental clues whose presence signifies potential threat or danger. Not only does the amygdala feed back to the visual cortex¹⁷, modulating even relatively early visual information processing^{18,19}, but as the present study suggests it might also influence the visual information that our eyes seek in the first place. This mechanism could be a component of the amygdala's role in the resolution of ambiguity in facial expressions²⁰ and the modulation of attention^{18,21,22}. Thus, we believe that the impaired fear recognition arising from damage to SM's amygdala is not due to a basic visuoperceptual inability to process information from the eyes, but is instead a failure by the amygdala to direct her visual system to seek out, fixate, pay attention to and make use of such information to identify emotions. This interpretation entails a revision of

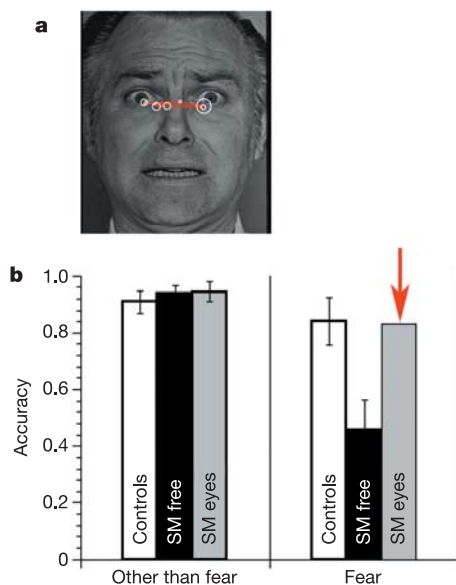


Figure 3 Instructed viewing of the eyes improves impaired fear recognition in SM. **a**, When instructed to fixate on the eyes in facial expressions of fear, SM is able to do so. **b**, Accuracy of emotion recognition ($\pm$ s.e.m.) for ten control subjects (white) and SM. Whereas SM's recognition of fear is impaired when allowed to look at the stimuli freely (SM free, black bars), her performance becomes normal relative to control subjects when instructed to fixate on the eyes (SM eyes, grey bar, red arrow). The impairment is specific to fear recognition (left panel shows mean recognition accuracy for all emotions other than fear).

Table 1 Mean accuracies in emotion recognition for SM and control subjects

Emotion	Controls	SM (free)	SM (eyes)
Happiness	1.00	1.00	1.00
Surprise	0.96	1.00	1.00
Anger	0.82	0.88	0.82
Disgust	0.76	0.85	0.90
Sadness	1.00	0.96	1.00
Fear	0.84	0.46	0.83

Subjects (SM and ten control subjects) were shown six different exemplars of each of six emotions using face stimuli¹³ identical to those used in prior studies¹, and were asked to identify the appropriate emotion by pushing a button. The experiment was conducted twice with controls and four times with SM: twice when she was allowed to look freely at the images (free), and twice when instructed to fixate on the eyes (eyes). The only significant difference between SM and control subjects is in her recognition of fear under the free viewing condition ($Z = -2.385$, $P < 0.01$, one-tailed t -test).

our previous conclusions¹ about the face processing abilities of SM: although she can generate a normal performance score on discrimination and recognition tasks for some emotions (such as happiness), her use of visual information is abnormal for all facial emotions, not only fear.

Our study is in line with recent findings that the amygdala participates in processing information about the eye region of faces^{6,23,24}. Such a functional specialization might account for the role of the amygdala in processing emotions related to behavioural withdrawal²⁵, fear²⁶, threat or danger^{3,7}. A strategy of directing one's own gaze onto the eyes of others would serve to seek out potential sources of salient social information²⁷, and it seems plausible that other impairments in social judgement resulting from bilateral amygdala damage⁹ could be attributed, at least in part, to the same mechanism. It is intriguing to consider the possibility that disorders such as autism, which also features impaired fixations to the features of faces^{28,29} and impaired processing of emotion from faces³⁰, might benefit from instructed viewing as we found in SM. □

Methods

Subjects

We tested subject SM, a 38-yr-old woman with bilateral amygdala damage, 30 neurologically normal females of comparable mean age (37.5 yr, s.d. = 3) and 13 neurological subjects with focal, unilateral amygdala damage due to temporal lobectomy (five subjects with right lobectomy (three females, two males) and eight subjects with left lobectomy (three females, five males), with a mean age of 37.4 yr, s.d. = 12). SM participated in all experiments, and control subjects participated as specified below. All subjects had normal or corrected-to-normal visual acuity, normal basic visuo-perception (for example, from the Benton facial recognition task) and IQ in the normal range.

Bubbles task

SM, along with ten normal control subjects and all 13 subjects with unilateral amygdala damage, were seated 1 m in front of a 17-inch LCD display in a dimly lit room. Images (5.72° × 5.72°) were shown at the centre of the screen one at a time with no time limit, until the subject pushed one of two buttons required for the discrimination task: either a discrimination between fear and happiness, or between male and female. Each block of trials consisted of one discrimination task. Faces were drawn randomly from the four exemplars shown in Supplementary Fig. 1 (see Supplementary Information for construction of stimuli), and sparsely sampled in the two-dimensional image plane and in five spatial frequency bands as described in detail elsewhere^{10,11} (see Supplementary Fig. 1). Gaussian bubbles were adjusted to maintain 75% correct discrimination performance for each subject. The emotion discrimination task consisted on average of 2,970 trials (3,072 trials for SM), and the gender task consisted on average of 2,048 trials. These were broken down into multiple sessions, for a cumulative testing time of 6–10 h per subject.

Faces with eyes erased

Twelve normal controls and SM were shown the same faces as in the eye-tracking tasks below, in two blocks separated by approximately 1 h. In the first block the eye region of the faces was replaced with a grey rectangle and in the second block the same faces were shown with the eyes present. Subjects were asked to pick one of the six emotion labels that best matched the stimulus.

Eye-tracking tasks

Eighteen normal control subjects and SM participated in this experiment. One experiment (ten subjects) consisted of five blocks of the conditions listed below, where each condition contained 39 face images from the Ekman data set, identical to those used in other studies¹ (six images of each of the six basic emotions and three neutral images). All stimuli were preceded by a 2 s interstimulus interval consisting of a grey screen isoluminant with the face stimuli, followed by a 1 s central fixation cross, followed by the face image for 5 s. The five blocks were presented in fixed order for all subjects as follows: (1) passive viewing; (2) judging the emotion, where subjects were asked to push one of six buttons corresponding to the emotion labels; (3) judging the gender, where subjects were asked to push one of two buttons; (4) judging the emotion (same as block 2); and (5) passive viewing (same as block 1).

A second experiment was run with SM and eight control subjects, using identical conditions to block 2 above, but with the fixation cross located in a position coincident with either the left or the right eye in each face (two blocks run for each subject).

Eye movements were measured with an infrared pupil-centred corneal-reflection eye tracker (Applied Science Laboratories, Model 504). The criterion for identifying a fixation was that the viewer's gaze drifted by less than 1° within 100 ms.

Data analysis

For the two-alternative discrimination task, we performed multiple linear regression using the gaussian bubble parameters (x and y coordinates specifying the bubble location on the two-dimensional face image, and a third parameter specifying spatial frequency scale) and the subject's accuracy throughout the task. This in effect yielded a three-dimensional

regression coefficient volume, which was transformed into a Z-score volume. This Z-score volume was visualized by assigning a threshold at $P < 0.05$.

For the eye-tracking tasks, we drew a box around the eye region of each stimulus (Fig. 2a) and counted the number of fixations made in this region during the entire 5-s presentation of the stimulus. This was then expressed as a proportion of the total number of fixations made on the entire face.

Received 7 September; accepted 7 October 2004; doi:10.1038/nature03086.

- Adolphs, R., Tranel, D., Damasio, H. & Damasio, A. Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature* **372**, 669–672 (1994).
- Young, A. W. *et al.* Face processing impairments after amygdalotomy. *Brain* **118**, 15–24 (1995).
- Adolphs, R. *et al.* Recognition of facial emotion in nine subjects with bilateral amygdala damage. *Neuropsychologia* **37**, 1111–1117 (1999).
- Schmolk, H. & Squire, L. R. Impaired perception of facial emotions following bilateral damage to the anterior temporal lobe. *Neuropsychology* **15**, 30–38 (2001).
- Morris, J. S. *et al.* A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature* **383**, 812–815 (1996).
- Adams, R. B., Gordon, H. L., Baird, A. A., Ambady, N. & Kleck, R. E. Effects of gaze on amygdala sensitivity to anger and fear faces. *Science* **300**, 1536 (2003).
- Adolphs, R. & Tranel, D. In *The Amygdala. A Functional Analysis* (ed. Aggleton, J. P.) 587–630 (Oxford Univ. Press, New York, 2000).
- Bechara, A. *et al.* Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science* **269**, 1115–1118 (1995).
- Adolphs, R., Tranel, D. & Damasio, A. R. The human amygdala in social judgment. *Nature* **393**, 470–474 (1998).
- Gosselin, F. & Schyns, P. G. Bubbles: a technique to reveal the use of information in recognition. *Vision Res.* **41**, 2261–2271 (2001).
- Schyns, P. G., Bonnar, L. & Gosselin, F. Show me the features! Understanding recognition from the use of visual information. *Psychol. Sci.* **13**, 402–409 (2002).
- Smith, M. L., Cottrell, G. W., Gosselin, F. & Schyns, P. G. Transmitting and decoding facial expressions of emotion. *Psychol. Sci.* (in the press).
- Ekman, P. & Friesen, W. *Pictures of Facial Affect* (Consulting Psychologists Press, Palo Alto, California, 1976).
- Lundqvist, D. & Litton, J.-E. *The Karolinska Directed Faces* (Karolinska Institute, 1998).
- Haith, M. M., Bergman, T. & Moore, M. J. Eye contact and face scanning in early infancy. *Science* **198**, 853–855 (1977).
- Keating, C. & Keating, E. G. Visual scan patterns of rhesus monkeys viewing faces. *Perception* **11**, 211–219 (1982).
- Amaral, D. G. & Price, J. L. Amygdalo-cortical connections in the monkey (*Macaca fascicularis*). *J. Comp. Neurol.* **230**, 465–496 (1984).
- Anderson, A. K. & Phelps, E. A. Lesions of the human amygdala impair enhanced perception of emotionally salient events. *Nature* **411**, 305–309 (2001).
- Vuilleumier, P., Richardson, M. P., Armony, J. L., Driver, J. & Dolan, R. J. Distant influences of amygdala lesion on visual cortical activation during emotional face processing. *Nature Neurosci.* **7**, 1271–1278 (2004).
- Whalen, P. J. Fear, vigilance, and ambiguity: initial neuroimaging studies of the human amygdala. *Curr. Dir. Psychol. Sci.* **7**, 177–187 (1999).
- Kapp, B. S., Whalen, P. J., Supple, W. F. & Pascoe, J. P. In *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction* (ed. Aggleton, J.) 254–269 (Wiley-Liss, New York, 1992).
- Holland, P. C. & Gallagher, M. Amygdala circuitry in attentional and representational processes. *Trends Cogn. Sci.* **3**, 65–73 (1999).
- Kawashima, R. *et al.* The human amygdala plays an important role in gaze monitoring. *Brain* **122**, 779–783 (1999).
- Morris, J. S., deBorja, M. & Dolan, R. J. Human amygdala responses to fearful eyes. *Neuroimage* **17**, 214–222 (2002).
- Anderson, A. K., Spencer, D. D., Fulbright, R. K. & Phelps, E. A. Contribution of the anteromedial temporal lobes to the evaluation of facial emotion. *Neuropsychology* **14**, 526–536 (2000).
- Calder, A. J. *et al.* Facial emotion recognition after bilateral amygdala damage: differentially severe impairment of fear. *Cogn. Neuropsychol.* **13**, 699–745 (1996).
- Emery, N. J. The eyes have it: the neuroethology, function and evolution of social gaze. *Neurosci. Biobehav. Rev.* **24**, 581–604 (2000).
- Pelphrey, K. A. *et al.* Visual scanning of faces in autism. *J. Autism Dev. Disord.* **32**, 249–261 (2002).
- Klin, A., Jones, W., Schultz, R., Volkmar, F. R. & Cohen, D. Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Arch. Gen. Psychiatry* **59**, 809–816 (2002).
- Howard, M. A. *et al.* Convergent neuroanatomical and behavioral evidence of an amygdala hypothesis of autism. *Neuroreport* **11**, 2931–2935 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank all participants for their time and effort; N. J. Cohen and members of his laboratory; D. Hannula and J. Gibbons, for help with the administration and analysis of eye movement data; H. Damasio and M. Koenigs for comments on the manuscript; B. Lewis and K. Scheer for assistance with testing subjects, and R. Henson for help in scheduling their visits. This work was supported by grants from NINDS and NIMH, and a Twenty-first Century Science grant from the James S. McDonnell Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.A. (radolphs@hss.caltech.edu).

β -Lactam antibiotics offer neuroprotection by increasing glutamate transporter expression

Jeffrey D. Rothstein^{1,2}, Sarjubhai Patel¹, Melissa R. Regan¹, Christine Haenggeli¹, Yanhua H. Huang², Dwight E. Bergles², Lin Jin¹, Margaret Dykes Hoberg¹, Svetlana Vidensky¹, Dorothy S. Chung¹, Shuy Vang Toan¹, Lucie I. Bruijn³, Zao-zhong Su⁴, Pankaj Gupta⁴ & Paul B. Fisher⁴

¹Department of Neurology, ²Department of Neuroscience, Johns Hopkins University, Baltimore, Maryland 21287, USA

³The ALS Association, Palm Harbor, Florida 34685, USA

⁴Columbia University Medical Center, College of Physicians and Surgeons, Department of Pathology, Neurosurgery and Urology, New York, New York 10032, USA

Glutamate is the principal excitatory neurotransmitter in the nervous system. Inactivation of synaptic glutamate is handled by the glutamate transporter GLT1 (also known as EAAT2; refs 1, 2), the physiologically dominant astroglial protein. In spite of its critical importance in normal and abnormal synaptic activity, no practical pharmaceutical can positively modulate this protein. Animal studies show that the protein is important for normal excitatory synaptic transmission, while its dysfunction is implicated in acute and chronic neurological disorders, including amyotrophic lateral sclerosis (ALS)³, stroke⁴, brain tumours⁵ and epilepsy⁶. Using a blinded screen of 1,040 FDA-approved drugs and nutritionals, we discovered that many β -lactam antibiotics are potent stimulators of GLT1 expression. Furthermore, this action appears to be mediated through increased transcription of the *GLT1* gene⁷. β -Lactams and various semi-synthetic derivatives are potent antibiotics that act to inhibit bacterial synthetic pathways⁸. When delivered to animals, the β -lactam ceftriaxone increased both brain expression of GLT1 and its biochemical and functional activity. Glutamate transporters are important in preventing glutamate neurotoxicity^{1,9–11}. Ceftriaxone was neuroprotective *in vitro* when used in models

of ischaemic injury and motor neuron degeneration, both based in part on glutamate toxicity¹¹. When used in an animal model of the fatal disease ALS, the drug delayed loss of neurons and muscle strength, and increased mouse survival. Thus these studies provide a class of potential neurotherapeutics that act to modulate the expression of glutamate neurotransmitter transporters via gene activation.

To identify compounds capable of increasing rodent GLT1 expression, a structurally diverse library of 1,040 FDA-approved drugs and nutritionals were individually added to organotypic spinal cord slice cultures prepared from postnatal day 9 rats (Fig. 1a). This approach mimics the cellular metabolism and cell-cell interactions present *in vivo*. All assays were conducted in a blinded fashion, and each drug (100 μ M, added biweekly) was studied in duplicate or triplicate (10–15 tissue samples per drug). After 5–7 days of drug treatment, tissue was harvested and immunoblotted for expression for GLT1 protein using GLT1 anti-peptide antibodies (Fig. 1b). Dibutylryl cyclic AMP, a *GLT1* gene activator, served as a positive control (Fig. 1b), while 0.1% DMSO controlled for drug solubilizer. Approximately 50–60 drugs per week were investigated. GLT1 protein was analysed by semiquantitative, semi-automated densitometry. Replicate variability was less than 10%. Over 20 compounds were capable of increasing GLT1 protein expression by more than twofold compared to untreated controls (Fig. 1c). Analysis of the top 2% of all hits revealed that a single class of compounds, β -lactam antibiotics, was overly represented. Fifteen different β -lactam antibiotics, including penicillin and its derivatives, as well as cephalosporin antibiotics, were highly active in stimulating GLT1 protein expression (Fig. 1d). Increased expression could be seen as early as 48 h after drug treatment. Validation of the most active drugs was confirmed by repeat treatment of spinal cord cultures with 10–100 μ M drug. The EC₅₀ for increasing GLT1 expression by a representative cephalosporin, ceftriaxone, was 3.5 μ M (Fig. 1e), which is comparable to the known central nervous system (CNS) levels attainable with therapy for meningitis (0.3–6 μ M)^{12,13}. Non- β -lactam antibiotics included in the screen had no effect on GLT1 protein expression, including kanamycin, fluconazole, minocycline, polymyxin and doxycycline.

To better understand the mechanism of action, the effect of the drugs on the GLT1 promoter was examined in cell lines from

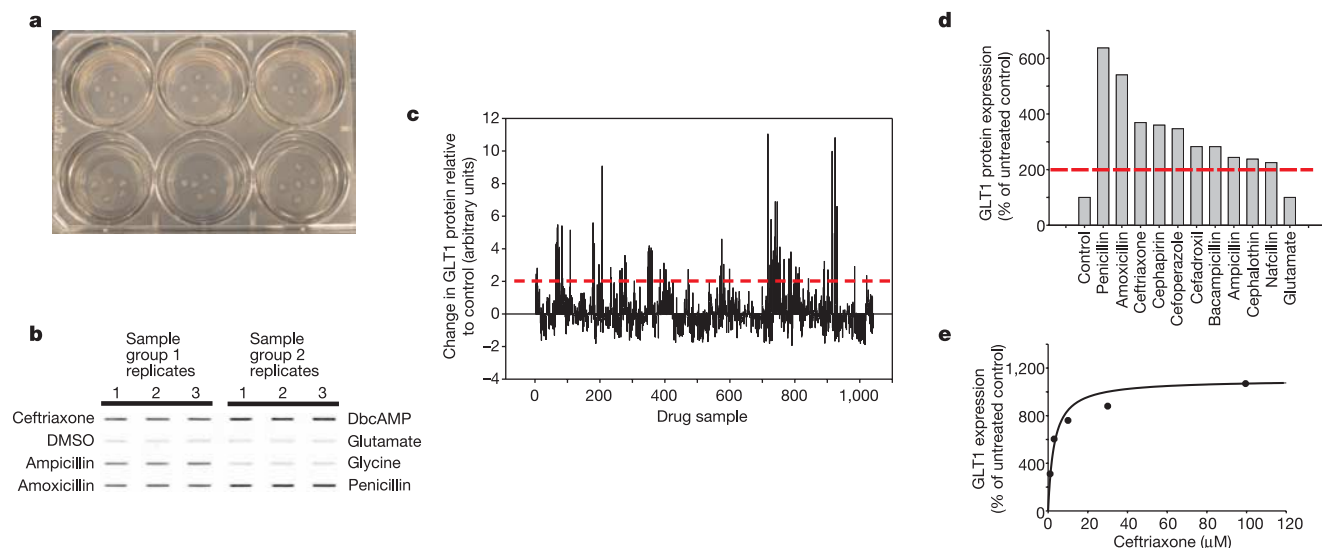


Figure 1 Screen of 1,040 FDA-approved drugs reveals β -lactam antibiotics as inducers of GLT1 protein expression. **a**, Rodent lumbar spinal cord cultures. **b**, Sample raw data slot blot of GLT1 protein in triplicate, including untreated tissue control, dibutylryl cyclic AMP positive control (dbcAMP), DMSO drug vehicle control, and various drugs (all shown

here at 10 μ M). **c**, Screening results for 1,040 sample compounds. Bar height reflects increased GLT1 protein expression relative to vehicle-treated controls. **d**, β -Lactam antibiotics were highly represented among the most potent compounds. **e**, Dose response analysis for ceftriaxone, revealing EC₅₀ of 3.5 μ M for GLT1 expression.

astrocytes and non-neuronal tissues. A 2.5-kilobase (kb) fragment of the human GLT1 promoter linked to firefly luciferase was transfected into human fetal astrocytes⁷, and used to screen the active compounds identified above. Similar results were also obtained using stable cell lines of human fetal astrocytes or COS7 cells transfected with a 2.7-kb GLT1 promoter fragment linked to both enhanced green fluorescent protein (eGFP) complementary DNA and firefly luciferase cDNA. As shown in Fig. 2a, the human GLT1 promoter fragment was significantly activated by ceftriaxone, amoxicillin and dibutyryl cyclic AMP, but not by the antibiotic vancomycin, amino acids glutamate and glycine, or the vehicle, DMSO. These effects were dose dependent, seen as early as 48 h after drug administration, and persisted for at least 7 days *in vitro* (Fig. 2a). Additional analysis of various cephalosporins (10 μ M) and β -lactams revealed prominent activity among the various agents (Fig. 2b), although the parent structure, cephalosporin C, was inactive in astroglial cell lines. No β -lactams were found that inhibited promoter activation.

As these compounds were capable of activating the promoter at concentrations known to be attainable in brain after parenteral administration (for example, 10–150 μ M)¹⁴, we further explored the *in vivo* biological activity of ceftriaxone in normal rats. After five to seven days of ceftriaxone therapy (200 mg per kg, i.p. daily, $n = 5$), animals were killed and brain tissue collected. Antibiotic treatment led to a threefold increase in GLT1 protein expression, and active splice variant GLT1b (ref. 15), as determined by semiquantitative immunoblots from hippocampus and spinal cord (Fig. 3a, b). This increase was persistent, and could also be observed after 3 months of treatment ($n = 10$). Conversely, the other molecular subtypes of glutamate transporters, including the astroglial protein GLAST and

the neuronal glutamate transporters EAAC1 and EAAT4, were unchanged after ceftriaxone administration (Fig. 3a, b).

GLT1 promoter activation was also observed *in vivo* (Fig. 3c, left panel). Chronic treatment of GLT1-BAC-eGFP promoter reporter mice with ceftriaxone produced an obvious increase in reporter expression in astroglial soma and processes throughout the hippocampal CA1 neuropil (Fig. 3c, right panel). Notably, in this brain region neuronal expression of the gene was not induced by drug (Fig. 3c, right panel). The effects of ceftriaxone appeared to be relatively specific, as the constitutive proteins actin (Fig. 3a) and superoxide dismutase 1 (SOD1, not shown), neuronal specific proteins neurofilament L and synaptophysin, and the astroglial protein glial fibrillary acid protein (GFAP), were unaffected by ceftriaxone therapy. Treatment with non- β -lactam antibiotics including vancomycin and minocycline had no effect on brain GLT1 levels.

Glutamate transporters are preferentially localized to astroglial membranes, although in some cases, increased protein expression is not always mirrored by concomitant membrane localization and functional activity¹⁶. However, cephalosporin therapy did increase biochemical glutamate transport, as measured by L-[³H]glutamate uptake into cortical membrane (Fig. 3d) or spinal cord (not shown) homogenates prepared from adult animals treated intraperitoneally for 7 days with drug. Similarly, after 7-day treatment, ceftriaxone increased GLT1-mediated L-[³H]-glutamate transport in a dose dependent fashion in cultured spinal cord slices (Fig. 3d). The increase in cell surface GLT1 was confirmed with cell membrane impermeant biotinylation reagent (Fig. 3e). Biotinylated GLT1 was increased on plasma membranes from mixed cortical neuron/astroglial cultures treated for 7 days with ceftriaxone. Finally, glutamate-transporter-associated currents tended to be larger in hippocampal astrocytes following 4–7 days of treatment of post-natal rat pups with ceftriaxone (Supplementary Data). Thus, *in vitro* and *in vivo* administration of ceftriaxone led to a threefold increase in protein levels and a comparable increase in GLT1 specific biochemical and electrophysiological transport. Penicillin treatment also increased biochemical transport (Fig. 3d), although its brain penetration is less, presumably accounting for the lower level of activity. Vancomycin was inactive in these functional assays.

Glutamate receptor antagonism has been extensively explored in acute and chronic neuroprotection, but no therapies exist to modulate glutamate-mediated injury via transporters. Genetic overexpression of transporters in transgenic mice and in engineered cell lines suggest that increasing the density of transporter in astroglia can be neuroprotective¹⁷. The level of neuroprotection may depend on the magnitude of overexpression. To determine if β -lactam antibiotics, ceftriaxone in particular, could be neuroprotective, we tested the compound in a series of *in vitro* and *in vivo* models. Treatment of cultured neurons with a low oxygen, low glucose condition, known as oxygen glucose deprivation (OGD), models the neuronal injury that can occur in ischaemic injury. In this model, one hour of OGD was lethal to cultured neurons, with toxicity known to involve excess glutamate¹⁸. However, when cultures are preconditioned 24 h before the lethal condition with transient OGD (5 min), there is a dramatic and robust resistance of neurons to cell death. This neuroprotection, referred to as ischaemic pre-conditioning, is due in part to increased expression of GLT1 (ref. 18), although some studies suggest these transporters could contribute to ischaemic injury². As shown in Fig. 4a, baseline neuronal death in the cultures was 14% (no treatment column, NT). Ceftriaxone (1 μ M), when added for 48 h to cultures, did not increase the baseline cell death (NT+ceftriaxone), but increased GLT1 protein levels (>25%; not shown) and transport. Cultures subject to 1 h OGD, without preconditioning, increased neuronal death to 50%. Ischaemic preconditioning OGD (5 min) applied 24 h before a one-hour OGD prevented neuronal injury. Importantly, 1 μ M ceftriaxone (or the β -lactam

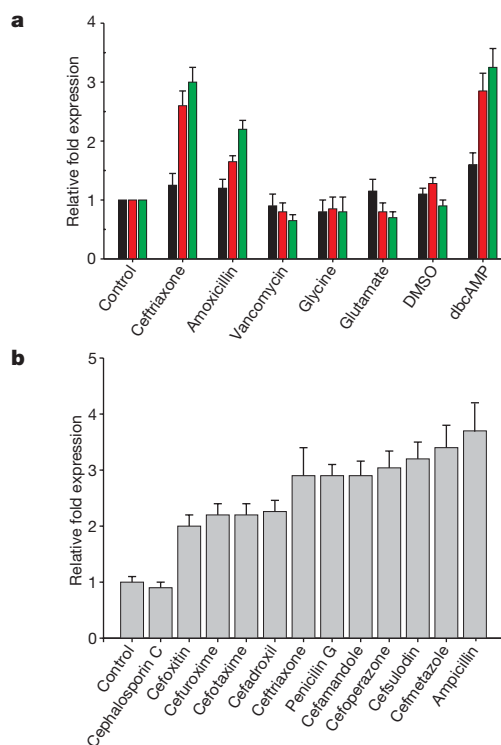


Figure 2 Promoter reporter analysis. β -Lactams activate human GLT1 promoter. **a**, In human fetal astrocytes transfected with the GLT1 promoter/luciferase reporter, β -lactam antibiotics at 0.1 μ M (black), 1 μ M (red) and 10 μ M (green), markedly activate the GLT1 promoter in a dose dependent manner, while controls such as glutamate and glycine have no effect. Dibutyryl cyclic AMP is a known GLT1 promoter activator. **b**, Closer analysis of cephalosporin antibiotics reveals consistent activation (10 μ M) by many, but not all, structural variants, while vancomycin had no effect. Data shown as mean+s.e.m.

cefuroxime, not shown), when added 48 h before 1 h OGD, was also protective, reducing the percentage of neuronal cell death from 50% to 20%—similar to ischaemic tolerance neuroprotection. Thus β -lactam pre-treatment appeared to prevent neuronal death in ischaemic tolerance.

Chronic blockade of glutamate transport in spinal cord organotypic cultures, with the non-specific transporter inhibitor *threo*- β -hydroxyaspartate (THA) or *DL*-*threo*- β -benzyloxyaspartate (TBOA)¹¹ leads to chronic increase in extracellular glutamate and subsequent slow death of motor neurons. To determine if ceftriaxone-induced GLT1 overexpression could be neuroprotective, we examined motor neuron degeneration in the organotypic spinal cord model. Organotypic cultures were prepared from lumbar spinal cords of 8–9-day-old rodent pups¹¹. No drugs were added for the first 7 days following culture preparation. Then ceftriaxone (1–100 μ M) was added with media changes, and after 7 more days, THA or TBOA were added at a concentration of 100 μ M, which produces chronic death of motor neurons. After 2–4 weeks, cultures were immunostained for neurofilament to quantify large ventral horn motor neurons. As shown in Fig. 4b, ceftriaxone treatment prevented motor neuron loss in a dose dependent manner. Similar neuroprotective results were seen with penicillin (not shown). As an additional control, organotypic spinal cord cultures prepared from GLT1-null mice were not protected from THA toxicity by ceftriaxone pre-treatment (Fig. 4b). Vancomycin was not protective.

To determine if ceftriaxone could alter neurodegeneration in a disease model that involves altered expression of glutamate transporters, we treated G93A SOD1 mice with drug. Studies have documented a contributory role for excess glutamate in this model, including neuroprotection by glutamate receptor blockade^{11,19–21}. Modest GLT1 overexpression can alter disease progression¹⁷. Guo *et al.*¹⁷ reported that a 1.5–2.3 fold increase in N-myc labelled human GLT1 expression in G93A SOD1 mice delayed disease onset as measured by grip strength ($\sim$ 14 days), but had no effect on other onset parameters such as weight loss and paralysis (3 days), and had no effect on survival. Initiating drug treatment in

this animal model around the time of clinical disease onset at, for example, loss of strength, most closely matches the use of human therapy, and could be more therapeutically relevant²². G93A SOD1 mice were treated daily with ceftriaxone (200 mg kg⁻¹ i.p.) starting at 12 weeks of age—approximately the time of clinical disease onset. Drug-treated animals ($n = 20$) and saline-injected controls ($n = 20$) were monitored daily for survival, and weekly for grip strength and body weight^{22,23}. As shown in Fig. 4c, d, ceftriaxone treatment significantly delayed loss of muscle strength and body weight. This effect was observed within 7 days after treatment, and persisted for 4–6 weeks. By 19 weeks of age, the strength preservation was lost. In a similar manner, the drug also increased overall survival of the mice by 10 days (ceftriaxone treated, 132 ± 2 days (all data with errors show mean $\pm$ s.e.m.); saline control, 122 ± 2 days; log rank, $\chi^2 = 7.8$, $P > 0.005$; Wilcoxon $\chi^2 = 7.5$, $P > 0.006$) (Fig. 4e). This effect is typical of drugs given relatively late in the life of G93A SOD1 mice, when the first clinical signs of disease are evident, and thus even a small effect may have clinical significance. When the same dose of drug was administered somewhat earlier, at 6 weeks of age, survival was also increased (ceftriaxone treated, 135 ± 2 days, $n = 20$; saline treated 122 ± 1.9 days, $n = 20$), although not significantly better than late delivery at 90 days of age. The lack of greater efficacy when given earlier would be consistent with the observation that the loss of GLT1 expression does not begin to occur until around 90 days in this model.

To determine if ceftriaxone altered cellular neurodegeneration *in vivo*, G93A mice were treated with ceftriaxone starting at 70 days of age. Two weeks of drug therapy lead to a significant prevention of motor neuron loss (Fig. 4h, i) and reduction of hypercellular gliosis compared to saline-treated control G93A mice. GLT1 expression decreases around the onset of clinical disease²⁴, yet ceftriaxone administration was able to increase endogenous GLT1 expression significantly in spinal cords from the chronically treated mice (Fig. 4f–h). The neuroprotection seen in this study was not likely to be due to the normal antibiotic properties of the drug, because ALS mice are not septic and do not have lung infections at

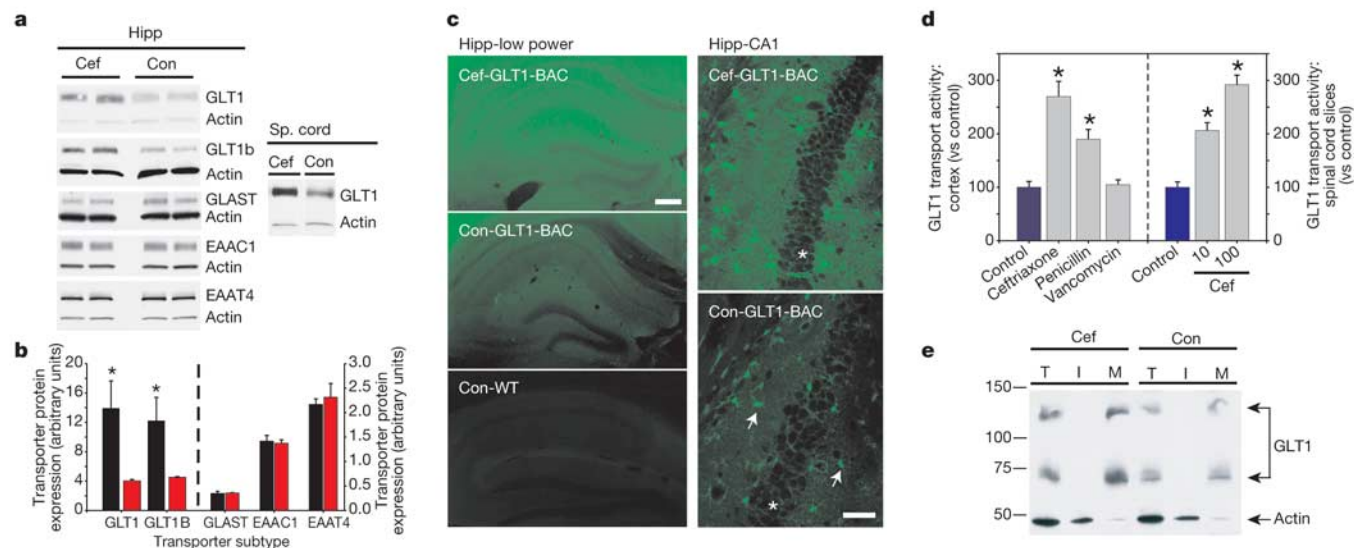


Figure 3 β -Lactam induces transporter promoter activation and protein expression *in vivo*. **a**, **b**, Ceftriaxone (black bar) induces expression of GLT1 and GLT1b protein, in hippocampus (Hipp) and spinal cord (sp. cord) (**a**, western blot) compared to saline control (**a**, control (con); **b**, red bar). Expression of the glutamate transporters GLAST, EAAC1 and EAAT4 were unaffected. **c**, Increased *in vivo* activation of the GLT1 promoter (**c**, left panels, low power light microscopy; scale bar, 200 μ m), using GLT1 BAC-eGFP promoter reporter mice, in hippocampal CA1 astrocytes (asterisks; **c**, right panel, confocal

microscopy; scale bar, 50 μ m) and neuropil but not CA1 neurons (arrows) from drug-treated mice, compared to untreated control mice. **d**, Ceftriaxone and penicillin administration increased ³H-glutamate transport in cortex homogenates from drug-treated mice (left panel) and treated spinal cord cultures (right panel). * $P < 0.05$ compared to untreated control. **e**, Immunoblots of total (T), intracellular (I) and biotinylated fractions (M) of mixed neuron/glia cortical cultures treated for 7 days with ceftriaxone (cef, 100 μ M). Molecular weight markers in kDa. In **b**, **d**, data are mean $\pm$ s.e.m.

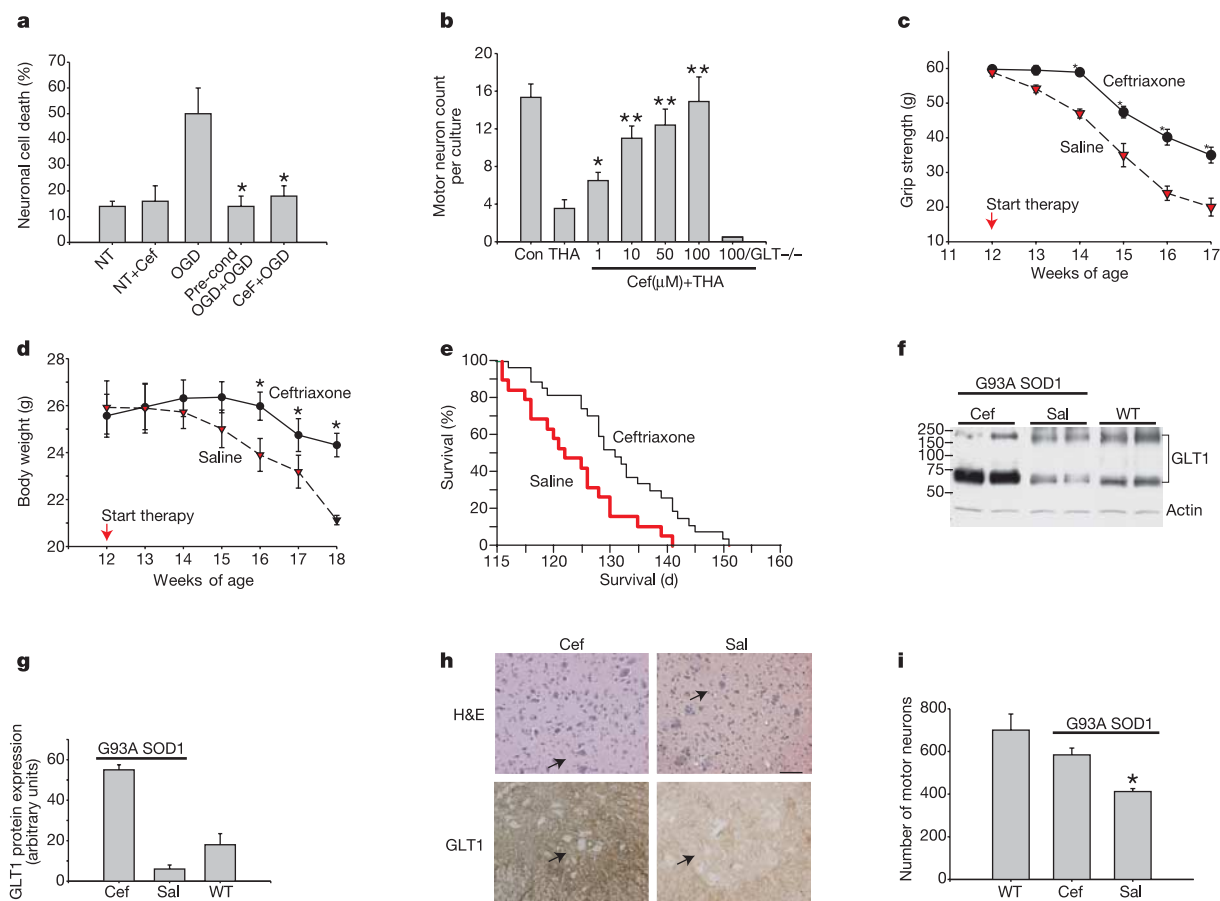


Figure 4 *In vitro* and *in vivo* neuroprotection by ceftriaxone. **a**, Oxygen glucose deprivation (OGD) of cultured cortical neurons was neurotoxic, but OGD preconditioning or ceftriaxone pre-treatment (1 μ M) was protective compared to no treatment (NT). **b**, Ceftriaxone (Cef) treatment of spinal cord cultures prevented *threo*-hydroxyaspartate (THA)-induced motor neuron loss but not in GLT1-null mouse (GLT1^{-/-}) tissue. * $P < 0.05$ or ** $P < 0.01$ versus untreated control. **c–e**, In G93A SOD1 ALS mice, ceftriaxone initiated at disease onset (red) delayed loss of muscle strength (**c**) and body weight (**d**) compared to saline treatment (black); ceftriaxone initiated at disease onset also increased survival (**e**). Spinal

cord GLT1 protein levels (**f**, **g**) and tissue expression (**h**) were markedly elevated in ceftriaxone (Cef)-treated ALS mice compared to saline (Sal)-treated ALS mice and untreated wild-type (WT) mice. Molecular weight markers in kDa. Two weeks of drug treatment (Cef) delayed loss of lumbar spinal motor neurons (**h**, **i**) compared to saline treatment in haematoxylin and eosin (H&E) stained tissue. Scale bar, 50 μ m. For panels **c**, **d** and **e**, $n = 20$ saline, $n = 20$ ceftriaxone group. * $P < 0.05$ versus ceftriaxone. Data in **a**, **b**, **g**, **i**, are mean $\pm$ s.e.m.; data in **c**, **d**, are mean $\pm$ s.e.m.

12–16 weeks of age—when prominent muscle strength effects were seen. In addition, the use of other CNS-penetrating antibiotics when given at this late stage (12 weeks old) do not prevent loss of muscle strength (for example, minocycline; L.I.B. and J.D.R., unpublished observations).

β -Lactam antibiotics, first identified with the discovery of penicillin in 1928, are now the most widely used antibiotics, and are one of the most important modern pharmaceuticals⁸. Notably, they have no substantial toxic CNS actions at normal antibacterial doses. Our studies document a new property of these antibiotics, and demonstrate that β -lactams can activate the gene for a neurotransmitter transporter. This is, to our knowledge, the first evidence of stimulatory pharmaceutical modulation of the glutamate transporter, and provides a new pathway for drug discovery and manipulation of glutamate transmission in disease. The mechanism of this over-expression appears to be activation of the genetic promoter for GLT1, although the pathway for promoter activation is as yet unknown. □

Methods

Screening assay and protein expression

Organotypic cultures were prepared from postnatal day 9 rat lumbar spinal cords¹¹. Slice cultures were maintained on Millicell-CM 30-mm inserts (5 slices per insert; Millipore,

PICM 03050) in 35-mm six-well plates (Falcon no. 3046) containing 1 ml growth media, without antibiotics, and maintained in a humidified atmosphere of 5% CO₂. After 7 days *in vitro*, cultures were treated with the NINDS Custom Collection (MicroSource Discovery) drugs at a concentration of 100 μ M for 7 days. Media and drugs were changed biweekly. GLT1 was quantified by slot blot (5 μ g protein per slot). Protein concentration in tissue sonicates was determined using Coomassie Plus protein assay (Pierce no. 1856210). GLT1, GLAST, EAAC1 and EAAT4 protein were detected using primary rabbit polyclonal rat anti-carboxy-terminal GLT-1 antibody followed by chemiluminescence (SuperSignal West Pico Chemiluminescent Substrate (Pierce no. 34080) detection (BioRad VersaDoc, Quantity One Discovery Series software, v4.3.0)). Twenty-one compounds were selected for retesting at 10–100 μ M to confirm hits (>300% of control). These compounds were also screened in a six-point dilution series with a maximum concentration of 300 μ M. These dilutions were created from freshly prepared 10 mM stocks in DMSO. Concentrations required to achieve 50% of the maximally achievable effect for each compound (EC₅₀) were calculated using SigmaPlot (Ver 9; Systat).

Human GLT1 promoter reporter assay

GLT1 promoter activity was studied in normal human fetal astrocytes seeded at 1×10^5 cells per 35-mm plate. Twenty-four hours after seeding, cells received the indicated compound at a final concentration of 1–10 μ M or were left untreated (control). Forty-eight hours later, the cells were transfected (calcium phosphate precipitation method⁷) with a pGL3/GLT1 luciferase reporter construct (5 μ g) plus a pSV β -galactosidase construct (1 μ g). In some cases, human fetal astrocytes or COS7 cell lines transfected with the GLT1 promoter (2.7 kb) luciferase/eGFP construct were used. After an additional 48 h, cell lysates were prepared and luciferase activity was determined using the Luciferase Assay System Kit (Promega, E1501) and luminescence determined using a luminometer (Turner Designs, TD20/20)⁷. Data presented is the average of three independent plates $\pm$ s.d.

GLT1 activity and immunoblotting

Levels of GLT1 protein were quantified by immunoblots³. Functional glutamate transport was measured by accumulation of ³H-glutamate in spinal cord slice or crude cortical synaptosomal membranes²⁵. Measurement of total glutamate uptake actually reflects the combined physiological activity of all transporter subtypes. GLT1 protein is uniquely sensitive to transport inhibition by dihydrokainate (DHK). To estimate the contribution of GLT1 to transport, aliquots of tissue homogenates were also incubated with 300 μ M DHK. Non-specific uptake was determined in the presence of 300 μ M *threo*- β -hydroxyaspartate (THA), at 0 °C and in sodium-free homogenates.

Generation of GLT1 BAC eGFP transgenic mouse

The BAC transgenic mice were generated as described previously²⁶ with a shuttle vector provided by N. Heintz. The BAC clone included approximately 45 kb upstream of the first GLT1 exon, the full GLT1 coding region (123 kb) and 24 kb downstream of the last exon. EGFP cDNA was inserted into the GLT1 start codon.

Oxygen glucose deprivation/ischaemic preconditioning

Primary cortical mixed neuronal-glial cell cultures were prepared from rodent fetal cortex (gestation day 14–16 CD1 mice) using the paradigm of ischaemic preconditioning¹⁸.

Motor neuron toxicity

Neuroprotection in spinal cord organotypic cultures prepared from postnatal day 8–9 wild-type rat or GLT1-null mouse tissue was performed as described previously¹¹. Ceftriaxone was added for 5–7 days before the addition of 100 μ M THA or DL-*threo*- β -benzyloxyspartate, (TBOA) 100 μ M. Surviving motor neurons were counted 2–3 weeks later by staining for phosphorylated neurofilaments (SMI-32).

G93A SOD1 mouse—disease onset and survival

Male transgenic mice expressing the human G93A SOD1 (B6SJL-TgN(SOD1-G93A)1Gur, high expressor) were bred with background-matched B6SJL wild-type females (Jackson Laboratories). The progeny were genotyped and used for subsequent studies. Experiments were conducted at Psychogenics (Hawthorne, New York) in accordance with protocols approved by the Johns Hopkins Animal Care and Use Committee. Mice were assessed by daily observation for survival, and by weekly weighing and testing of grip strength starting at 12 weeks of age^{22,23}. All experiments were performed blinded with coded syringes for injection.

Histology and motor neuron counts

Mice were perfused via cardiac infusion with 4% buffered paraformaldehyde and spinal cord post fixed with the same solution. The lumbar enlargement was collected, paraffin embedded, and serially sectioned at 14 μ m, for a total of 140 sections. Every seventh section was stained with haematoxylin and eosin, and examined at 20 $\times$ for motor neuron identification and counting²². Images were acquired using the Zeiss LSM 510 Meta confocal microscope (argon laser setting at 488 nm) with the operator blinded to treatment groups. All images were captured with the same gain, offset, pinhole diameter (2.53 Airy units), and scan speed (12.8 μ s with scan averaging set to 2). Z-series images were collected at 1.03 μ m intervals.

Statistics

Quantitative differences between *in vitro* and *in vivo* drug effects were analysed by analysis of variance (ANOVA) or Students *t*-test. Survival analysis was performed by Kaplan-Meier analysis. Software for statistics included Statview, and JMP 5.1 (SAS Software).

Received 11 July; accepted 4 November 2004; doi:10.1038/nature03180.

- Rothstein, J. D. *et al.* Knockout of glutamate transporters reveals a major role for astroglial transport in excitotoxicity and clearance of glutamate. *Neuron* **16**, 675–686 (1996).
- Danbolt, N. C. Glutamate uptake. *Prog. Neurobiol.* **65**, 1–105 (2001).
- Rothstein, J. D., Van Kammen, M., Levey, A. I., Martin, L. J. & Kuncel, R. W. Selective loss of glial glutamate transporter GLT-1 in amyotrophic lateral sclerosis. *Ann. Neurol.* **38**, 73–84 (1995).
- Rao, V. L. *et al.* Antisense knockdown of the glial glutamate transporter GLT-1, but not the neuronal glutamate transporter EAAC1, exacerbates transient focal cerebral ischemia-induced neuronal damage in rat brain. *J. Neurosci.* **21**, 1876–1883 (2001).
- Ye, Z. C., Rothstein, J. D. & Sontheimer, M. Compromised glutamate transport in human glioma cells: reduction-mislocalization of sodium-dependent glutamate transporters and enhanced activity of cystine-glutamate exchange. *J. Neurosci.* **19**, 10767–10777 (1999).
- Sepkuty, J. P. *et al.* A neuronal glutamate transporter contributes to neurotransmitter GABA synthesis and epilepsy. *J. Neurosci.* **22**, 6372–6379 (2002).
- Su, Z. Z. *et al.* Insights into glutamate transport regulation in human astrocytes: cloning of the promoter for excitatory amino acid transporter 2 (EAAT2). *Proc. Natl Acad. Sci. USA* **100**, 1955–1960 (2003).
- Goodman, L. S., Hardman, J. G., Limbird, L. E. & Gilman, A. G. *Goodman & Gilman's The Pharmacological Basis of Therapeutics* (McGraw-Hill Medical Pub. Division, New York, 2001).
- Tanaka, K. *et al.* Epilepsy and exacerbation of brain injury in mice lacking the glutamate transporter GLT-1. *Science* **276**, 1699–1702 (1997).
- Watake, K. *et al.* Motor discoordination and increased susceptibility to cerebellar injury in GLAST mutant mice. *Eur. J. Neurosci.* **10**, 976–988 (1998).
- Rothstein, J. D., Jin, L., Dykes-Hoberg, M. & Kuncel, R. W. Chronic inhibition of glutamate uptake produces a model of slow neurotoxicity. *Proc. Natl Acad. Sci. USA* **90**, 6591–6595 (1993).
- Chandrasekar, P., Rolston, K., Smith, B. & LeFrock, J. Diffusion of ceftriaxone into the cerebrospinal fluid of adults. *J. Antimicrob. Chemother.* **14**, 427–430 (1984).
- Nau, R. *et al.* Passage of cefotaxime and ceftriaxone into cerebrospinal fluid of patients with uninflamed meninges. *Antimicrob. Agents Chemother.* **37**, 1518–1524 (1993).
- Kazragis, R., Dever, L., Jorgensen, J. & Barbour, A. *In vivo* activities of ceftriaxone and vancomycin

against *Borrelia* spp. in the mouse brain and other sites. *Antimicrob. Agents Chemother.* **38**, 2632–2636 (1996).

- Chen, W. *et al.* Expression of a variant form of the glutamate transporter GLT1 in neuronal cultures and in neurons and astrocytes in the rat brain. *J. Neurosci.* **22**, 2142–2152 (2002).
- Schlag, B. D. *et al.* Regulation of the glial Na⁺-dependent glutamate transporters by cyclic AMP analogs and neurons. *Mol. Pharmacol.* **53**, 355–369 (1998).
- Guo, H. *et al.* Increased expression of the glial glutamate transporter EAAT2 modulates excitotoxicity and delays the onset but not the outcome of ALS in mice. *Hum. Mol. Genet.* **12**, 2519–2532 (2003).
- Romera, C. *et al.* *In vitro* ischemic tolerance involves upregulation of glutamate transport partly mediated by the TACE/ADAM17-tumor necrosis factor- α pathway. *J. Neurosci.* **24**, 1350–1357 (2004).
- Spalloni, A. *et al.* Cu/Zn-superoxide dismutase (GLY93 $\rightarrow$ ALA) mutation alters AMPA receptor subunit expression and function and potentiates kainate-mediated toxicity in motor neurons in culture. *Neurobiol. Dis.* **15**, 340–350 (2004).
- Canton, T. *et al.* RPR 119990, a novel alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid antagonist: synthesis, pharmacological properties, and activity in an animal model of amyotrophic lateral sclerosis. *J. Pharmacol. Exp. Ther.* **299**, 314–322 (2001).
- Rothstein, J. D. & Kuncel, R. W. Neuroprotective strategies in a model of chronic glutamate-mediated motor neuron toxicity. *J. Neurochem.* **65**, 643–651 (1995).
- Kaspar, B. K., Llado, J., Sherkat, N., Rothstein, J. D. & Gage, F. H. Retrograde viral delivery of IGF-1 prolongs survival in a mouse ALS model. *Science* **301**, 839–842 (2003).
- Drachman, D. B. *et al.* Cyclooxygenase 2 inhibition protects motor neurons and prolongs survival in a transgenic mouse model of ALS. *Ann. Neurol.* **52**, 771–778 (2002).
- Howland, D. S. *et al.* Focal loss of the glutamate transporter EAAT2 in a transgenic rat model of SOD1 mutant-mediated amyotrophic lateral sclerosis (ALS). *Proc. Natl Acad. Sci. USA* **99**, 1604–1609 (2002).
- Rothstein, J. D., Martin, L. J. & Kuncel, R. W. Decreased glutamate transport by the brain and spinal cord in amyotrophic lateral sclerosis. *N. Engl. J. Med.* **326**, 1464–1468 (1992).
- Gong, S., Yang, X. W., Li, C. & Heintz, N. Highly efficient modification of bacterial artificial chromosomes (BACs) using novel shuttle vectors containing the R6Kgamma origin of replication. *Genome Res.* **12**, 1992–1998 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to J. Lee and C. Cocci for technical assistance; K. Tanaka for GLT1-null mice; C. Leahy for ALS mouse studies; and J. Heemskerk for initiating the project, discussions and encouragement. G93A SOD1 mice were provided by Project ALS. The work was supported by the NIH, the Muscular Dystrophy Association and The Robert Packard Center for ALS Research at Johns Hopkins.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to J.D.R. (jrothste@jhmi.edu).

Nucleolar proteome dynamics

Jens S. Andersen¹†, Yun W. Lam²†, Anthony K. L. Leung^{2*}, Shao-En Ong¹, Carol E. Lyon², Angus I. Lamond² & Matthias Mann¹

¹Department of Biochemistry and Molecular Biology, Campusvej 55, DK-5230 Odense M, Denmark

²Wellcome Trust Biocentre, MSI/WTB Complex, University of Dundee, Dundee DD1 4HN, UK

* Present address: Center for Cancer Research, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

† These authors contributed equally to this work

The nucleolus is a key organelle that coordinates the synthesis and assembly of ribosomal subunits and forms in the nucleus around the repeated ribosomal gene clusters. Because the production of ribosomes is a major metabolic activity, the function of the nucleolus is tightly linked to cell growth and proliferation, and recent data suggest that the nucleolus also plays an important role in cell-cycle regulation, senescence and stress responses^{1–4}. Here, using mass-spectrometry-based organellar proteomics and stable isotope labelling⁵, we perform a quantitative analysis of the proteome of human nucleoli. *In vivo* fluorescent imaging techniques are directly compared to endogenous protein changes measured by proteomics. We characterize the flux of 489 endogenous nucleolar proteins in response to three different metabolic inhibitors that each affect nucleolar morphology.

Proteins that are stably associated, such as RNA polymerase I subunits and small nuclear ribonucleoprotein particle complexes, exit from or accumulate in the nucleolus with similar kinetics, whereas protein components of the large and small ribosomal subunits leave the nucleolus with markedly different kinetics. The data establish a quantitative proteomic approach for the temporal characterization of protein flux through cellular organelles and demonstrate that the nucleolar proteome changes significantly over time in response to changes in cellular growth conditions.

HeLa cell nucleoli were isolated in high purity by density gradient fractionation⁶. Analysis of isolated nucleoli by electron microscopy shows that they remain intact and preserve the internal morphology seen *in situ* (Supplementary Fig. 1a, b). The preparations are homogeneous with virtually all particles visible by electron microscopy corresponding to nucleoli (Supplementary Fig. 1c and other data not shown). The isolated nucleoli retain transcriptional activity, as determined by 5-bromo-UTP (Br-UTP) incorporation (Supplementary Fig. 1d). The inhibition of transcription caused by actinomycin D treatment was confirmed by 5-fluorouridine incorporation analysis of control and actinomycin D-treated HeLa cells (Supplementary Fig. 1e, f).

Protein mixtures from individual one-dimensional gel slices were in gel digested with either trypsin or endoproteinase Lys-C. The resulting peptide mixtures were analysed in several runs of nano-scale liquid chromatography—tandem mass spectrometry (LC MS/MS) on an ion trap—Fourier Transform mass spectrometer, capable of very high mass accuracy and of sequencing several peptides per second (see Methods). To reduce false positives caused by random impurities, we independently prepared nucleoli in our two laboratories and only included proteins identified in both

preparations in the final proteome. Tandem mass spectra were searched in the human sequence database, considering only peptides conforming to full trypsin or Lys-C specificity and whose mass matched the calculated mass within 3 p.p.m. (see Methods). Approximately 11,130 unique peptide sequences were unambiguously matched to human genes with an average mass accuracy of 0.7 p.p.m. (see Supplementary Table 1). These peptides were used to identify proteins in the purified nucleoli with high stringency, requiring at least two high-scoring peptides per protein (see Supplementary Table 1). Under these conditions, we estimate a false positive rate for protein identification below 0.1% (see Methods).

We compared this expanded nucleolar proteome with published information on proteins localized to the nucleolus in budding yeast from a recent large-scale study of yeast protein localization using green fluorescent protein (GFP) fusion proteins⁷. Out of the 142 yeast nucleolar proteins that have at least one human homologue, 124 are found in the updated nucleolar proteome (87%). In our previous study⁸, ~58% of the yeast homologues were present in the proteome of 271 human factors. A comparison of the increasing fraction of the human homologues of yeast nucleolar proteins found in each separate determination of the HeLa cell nucleolar proteome during the past two years suggests that detection levels are approaching saturation coverage, at least with the technology currently available (Fig. 1b). The data indicate that approximately 90% of the yeast nucleolar proteins with human homologues are also nucleolar components in HeLa cells and that the nucleolus is highly conserved throughout the eukaryotic kingdom.

Werner's syndrome protein, which was not sequenced at all in our previous study, was here detected with six different peptides, demonstrating that the combination of increased sensitivity,

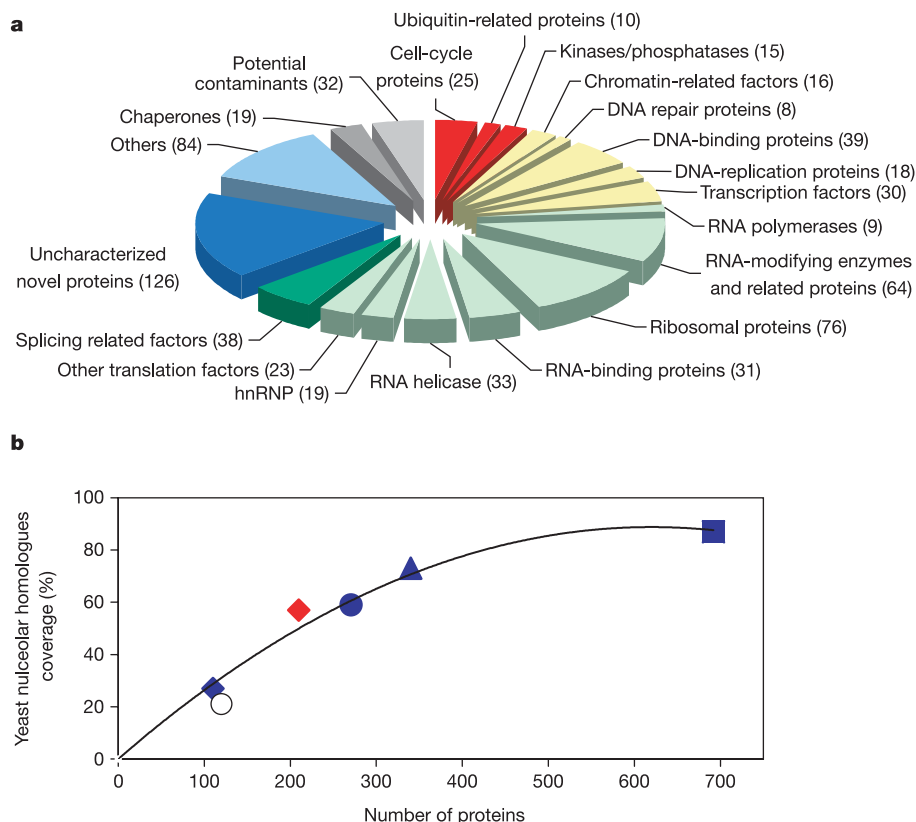


Figure 1 The nucleolar proteome. **a**, Classification of nucleolar proteins by functional category. Numbers indicate identified proteins in each category. **b**, Coverage of human homologues of yeast nucleolar proteins in different data sets. Open circle, proteins

annotated as nucleolar in the literature; blue diamond, MALDI study in ref. 8; red diamond, LC MS/MS study from ref. 30; blue circle, all proteins from ref. 8; blue triangle, LC MS/MS with QSTAR (this study); blue square, LTQ-FT data (this study).

resolution and peptide sequencing speed achieved much greater sequencing depth. Regulatory proteins in the nucleolar proteome included casein kinase II and the phosphatase PP1, which may have roles in nucleolar regulation^{9,10}, and protein kinases CDK2, CDK7, CDK9, CDC2L5, Cip1/p21 interacting protein and Aurora B, which are involved in controlling cell-cycle progression. We also observed key regulators of p53, such as p14ARF, and the Ser/Thr protein kinases VRK1 and ATM, which were reported to be sequestered in the nucleolus. Nonetheless, there remain some proteins previously reported to reside in the nucleolus that were not detected, such as the transcription factor RRN3 (ref. 11) and gemin 4 (ref. 12), which is a component of the survival of motor neuron (SMN) complex. Possible reasons for their continued absence from the observed proteome could include that they are present in exceedingly small amounts, or are not associated sufficiently stably with nucleoli to be isolated using the purification methods available. The data set from all our nucleolar mass spectrometry analyses defines an updated group of 692 proteins that reproducibly copurify with human nucleoli (also available in a searchable online database at <http://www.lamondlab.com/NOPdb/>). Bioinformatic classification (Fig. 1a) demonstrates functional diversity of the nucleolar proteome and the presence of approximately one-third of proteins with no previous functional information. Note that many proteins that stably copurify with nucleoli are also present at other cellular locations and some only accumulate transiently in nucleoli. The proteins identified included a minor fraction of potential contaminants, which have been included for completeness but are marked in Supplementary Table 1.

Unlike cytosolic organelles, nuclear bodies are not membrane bound and the principles of their organization and assembly are not well understood¹³. Recent light microscopy studies, analysing nuclear fluorescent fusion proteins using techniques such as fluorescence recovery after photobleaching (FRAP) or fluorescence loss in photobleaching (FLIP), have revealed the high mobility of many nuclear factors and their often rapid exchange between nuclear bodies and the surrounding nucleoplasm^{14,15}. Temporal studies have uncovered evidence for maturation pathways in which nuclear factors can transit between different nuclear bodies in a defined sequence¹⁶. To evaluate in a comprehensive and quantitative approach how dynamic the nucleolar proteome may be, we performed a series of proteomic studies on nucleoli isolated either from control cells, or from cells treated with drugs that inhibit transcription or protein degradation. We used stable-isotope labelling by amino acids in cell culture (SILAC)¹⁷ to characterize the response of the nucleolar proteome to transcription inhibition (Fig. 2). A HeLa cell line was metabolically labelled with either normal arginine ($^{12}\text{C}_6^{14}\text{N}_4\text{-Arg}$, termed here Arg0), carbon substituted arginine ($^{13}\text{C}_6^{14}\text{N}_4\text{-Arg}$, termed here Arg6) or carbon plus nitrogen substituted arginine ($^{13}\text{C}_6^{15}\text{N}_4\text{-Arg}$, termed here Arg10) respectively. This 'triple encoding' procedure allows three cell states to be measured in one experiment¹⁸. The cells are identical in all respects except that peptides derived after proteolytic digestion of the proteins can be distinguished in the mass spectrometer by their offsets of either zero, six or ten mass units.

The three HeLa cell populations were treated with actinomycin D at a final concentration of $1\text{ }\mu\text{g ml}^{-1}$, which inhibits transcription by RNA polymerase I, II and III¹⁹ for different lengths of time. A decrease in RNA synthesis levels, assayed by Br-UTP incorporation, was detected within 15–30 min, reflecting the time required for drug uptake by the cells (data not shown). Next, equal amounts of cells from each time point were mixed and nucleoli isolated directly from this mixed cell pool (Fig. 2a). Nucleolar proteins were fractionated and analysed by LC MS/MS as before. Because every arginine-containing tryptic peptide occurs in three isotopic forms, the intensities of these three mass spectrometry peaks directly reveal the relative ratios of the corresponding protein in the nucleolus at each of the three time points. Typically, several peptides per protein

are quantified (see Methods and Supplementary Table 2). Figure 2b, c shows representative mass spectra for peptides derived from p68. The increased peak heights for the 20 (Arg6) and 80 (Arg10) minute time points show that p68 is recruited to the nucleolus (Fig. 2d).

The kinetic experiments were performed using the HeLa^{YFP-p68} cell line⁸, allowing a direct comparison of changes in the level of YFP-p68 determined either by this proteomic analysis, or *in vivo* by digital fluorescence microscopy (Fig. 3). Western blotting analysis showed that the ratio between the yellow fluorescent protein (YFP)-tagged and untagged, endogenous p68 in isolated nucleoli remained the same during actinomycin D treatment and that both similarly increased in nucleoli when transcription was blocked (Fig. 3a). We monitored by microscopy the YFP-p68 fluorescence signal in nucleoli of live HeLa^{YFP-p68} cells following actinomycin D treatment and compared these *in vivo* data with the SILAC measurements of YFP-p68 in isolated nucleoli after actinomycin D treatment (Fig. 3b, c). Similar relative changes in nucleolar p68 levels were measured by both techniques (Fig. 3c and other data not shown). There was no overall change in the total cellular YFP-p68 fluorescence level following actinomycin D treatment, demonstrating that the increased nucleolar p68 fluorescence results from the intranuclear redistribution of p68 (Fig. 3c). A similar close agreement of fluorescence and SILAC data was also seen for two additional cell lines analysed that stably express either YFP-tagged fibrillarin (Fig. 3d) or YFP-tagged ribosomal protein L27 (Fig. 3e). Although it is possible that some changes in protein levels could occur after blocking transcription, we don't expect this to have a major effect on nucleolar proteome dynamics during the 2 h time course of transcription inhibition tested because most mammalian messenger

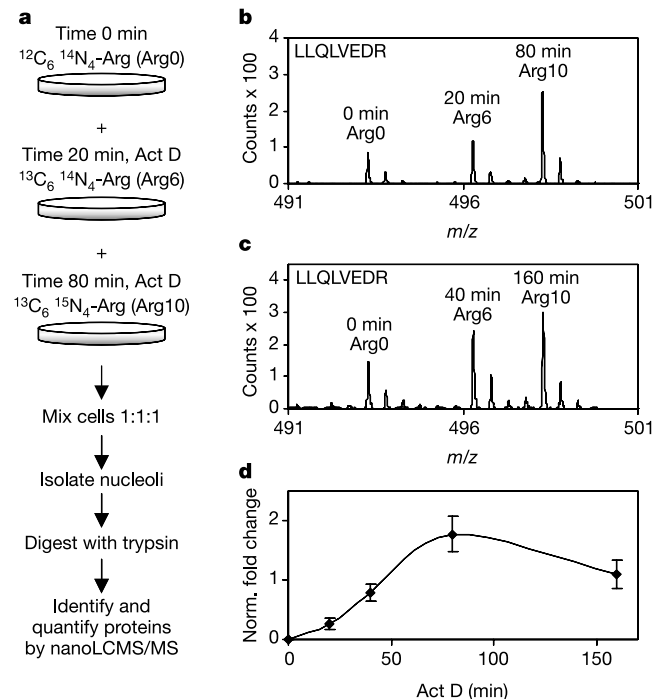


Figure 2 Determination of nucleolar protein dynamics. **a**, The proteomes in three cell populations are encoded by incorporation of stable isotope derivatives of arginine (SILAC method). Cells are metabolically labelled with Arg0, Arg6 and Arg10 for at least five cell doublings and are then treated for 0, 20 and 80 min, respectively. Cells are mixed and nucleoli purified and analysed by mass spectrometry. The analysis is repeated with a common zero point and additional time points of transcription inhibition to achieve higher time resolution. **b**, **c**, Spectra of peptides of p68, indicating increasing amounts p68 recruited to the nucleolus. **d**, Dynamic profiles of p68. Y axis is in units of normalized fold change of p68. Error bars are s.d. from several p68 peptides.

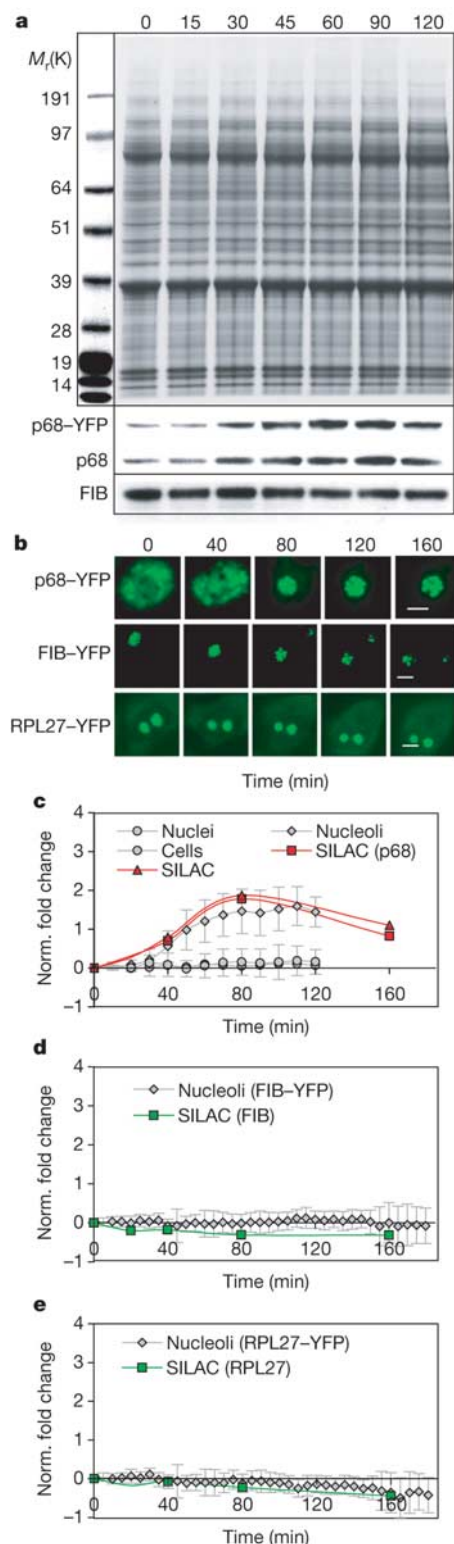


Figure 3 Comparison of different methods to measure nucleolar protein dynamics. **a**, Proteins from nucleoli isolated from HeLa YFP-p68 cells treated with actinomycin D (1 $\mu\text{g ml}^{-1}$) for 0–120 min were probed with antibodies (see Methods). **b–e**, Live HeLa cells stably expressing YFP-p68, YFP-fibrillarin (FIB) and YFP-RPL27 were treated with actinomycin D (1 $\mu\text{g ml}^{-1}$) and imaged for at least 120 min. The changes in the intranucleolar YFP fluorescence signals (grey curves) were compared to the changes in levels of the corresponding endogenous proteins detected in isolated nucleoli by SILAC (red or green curves); the YFP-p68 protein was also detected by SILAC. Scale bars, 3 μm . Error bars are s.d. from fluorescence measurements on five individual cells in each case.

RNAs have a half-life greater than 6 h (ref. 20). As illustrated for all three proteins tested, the changes in protein levels in nucleoli result primarily from redistribution. In addition, it was possible to resolve separate peptides from the YFP-tagged p68 and the endogenous untagged p68 proteins, allowing a comparison of their kinetic profiles in response to actinomycin D (Fig. 3c). Importantly, both show a similar kinetic response, indicating that the presence of the YFP tag does not significantly alter the dynamic behaviour of p68. Thus, the combination of mass spectrometry and fluorescence kinetic data measurements offers a powerful new way to validate these cell lines as model systems for *in vivo* studies of nucleolar dynamics.

Within the detected nucleolar proteome, 489 proteins produced peptides that allowed SILAC analysis. The relative levels of all 489 factors were therefore quantified in two large-scale experiments, measured at five or nine separate time points after inhibiting transcription with actinomycin D (Supplementary Table 2). The data show good reproducibility between the two experiments (Fig. 4 and see below). We observed a wide range of responses to actinomycin D treatment for different proteins (Fig. 4). The steady-state levels of many nucleolar proteins decreased to various extents on actinomycin D treatment. Factors known to be depleted from the nucleolus upon transcription inhibition (for example, PTB, Ki-67, GU and NOH61) were also found to decrease in our data set (Fig. 4b). Factors depleted from nucleoli after actinomycin D treatment included ribosomal proteins, RNA processing factors, exosome components and RNA polymerase I (Fig. 4a). This probably reflects the export of assembled ribosomal subunits taking place without ongoing RNA polymerase I activity to synthesize new ribosomal RNA substrate, and suggests that the association of many components with the nucleolus is dependent on ribosome subunit synthesis. A subset of proteins accumulated within nucleoli after transcription had been arrested, including all 11 proteins we identified previously to accumulate in nucleoli after actinomycin D treatment⁸. Remarkably, the level of some proteins increased up to tenfold (Fig. 4a). This result suggests that the nucleolus is not a simple ribosome synthesis machine that progressively breaks down in the absence of transcription. Instead, transcription inhibition leads to a more subtle redistribution of nuclear proteins, and the partition of proteins between the nucleolus and nucleoplasm may reflect the general physiological status of the cell.

Different nucleolar factors show major differences in their kinetics, of either decrease or accumulation. For example, separate nucleolar members of the DEAD box helicase family, which are likely to have distinct and non-overlapping functions in the nucleolus, respond to transcription inhibition with different kinetics (Fig. 4c). In contrast, separate subunits of RNA polymerase I, which physically associate in a stable protein complex, each leave nucleoli at an almost identical rate (Fig. 4d). These data are consistent with RNA polymerase I proteins moving coordinately, either as part of stable complexes or through the actions of common receptors. Similarly, separate small nuclear ribonucleoprotein particle (snRNP) proteins accumulate within nucleoli at identical rates, consistent with their association in a stable RNP complex. Multiple protein components of the exosome and the RNase P complex each also showed coordinate kinetics (Fig. 4e). Whereas multiple components appear to accumulate within the nucleolus only in the presence of rRNA substrate, the fact that many other proteins either remain at similar levels, or even increase within nucleoli after transcription inhibition, indicates that not all resident nucleolar proteins are dependent upon rRNA substrate synthesis. Thus the data do not support the view that the nucleolus is a transient structure formed only as a result of ongoing rRNA synthesis and indicate that at least a core of the nucleolar proteins remain associated in the absence of nascent rRNA transcription and ribosomal subunit assembly. The protein components of the large and small ribosomal subunits leave nucleoli with markedly different

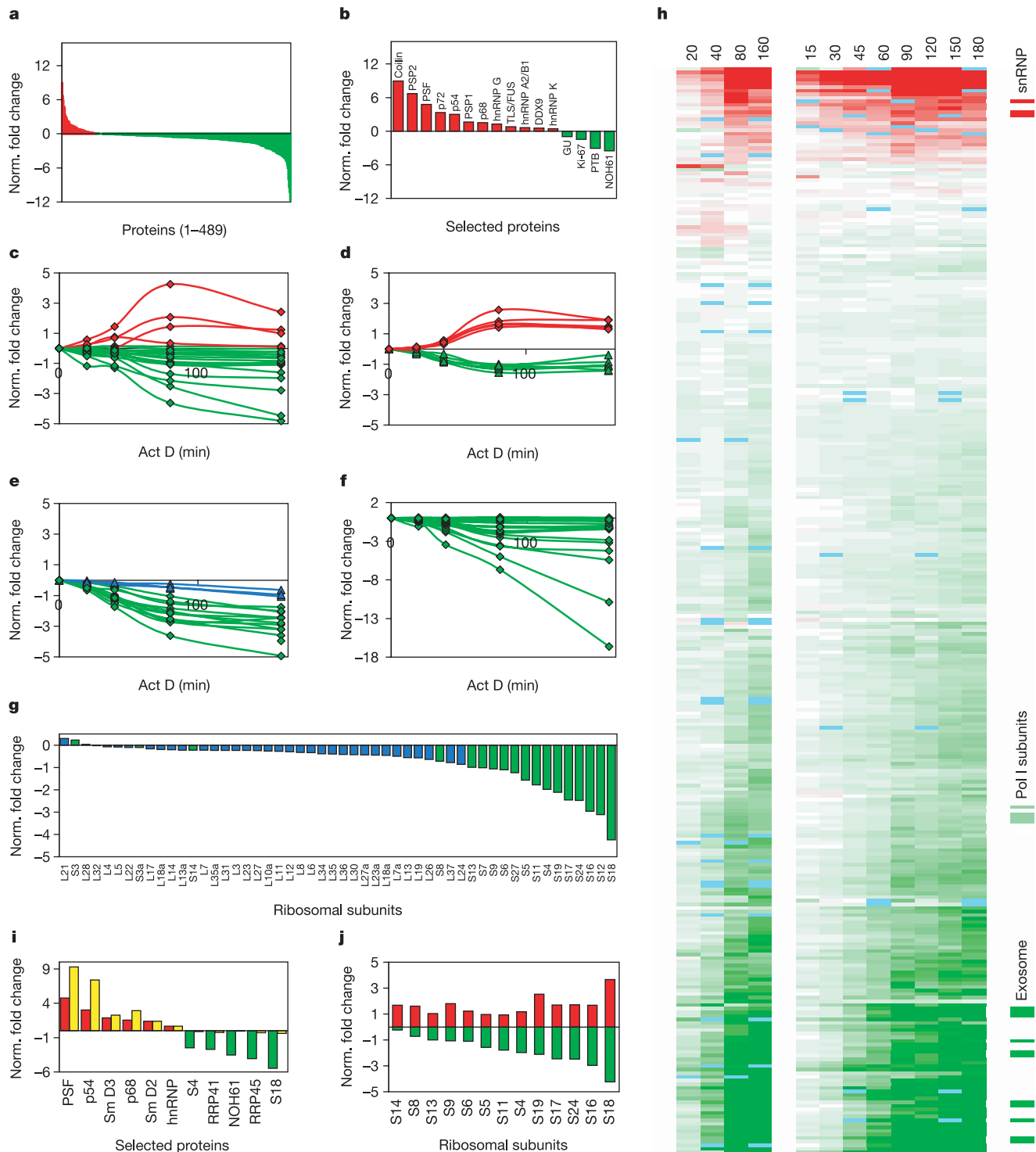


Figure 4 Dynamic profiles of nucleolar proteins. Red indicates recruited proteins and green indicates depleted proteins. **a**, All proteins showing change from first to last time points. **b**, Proteins known from the literature to be recruited to or depleted from nucleoli upon treatment with actinomycin D (Act D). **c**, Different kinetic profiles for different DEAD box proteins (top curve to bottom curve, see Supplementary Table 2: BAT1, CHD4, DDX10, DDX17, DDX18, DDX21, DDX24, DDX27, DDX31, DDX3X, DDX48, DDX49, DDX5, DDX50, DDX51, DDX52, DDX54, DDX56, DHX33, DHX37, MTR4, RUVBL2). **d**, Dynamic profile for polymerase I subunits (green; POLR1C, PAF53, POLR1B, POLR1A, POLR1D, TTF1, UBTf) and snRNP proteins (red; SNRPB, SNRPA, SNRPD2, SNRPD3, SNRPF). **e**, Dynamic profiles for subunits of the exosome (green; RRP42, RRP46, RRP40, RRP4, RRP43, RRP41, RRP45, CSL4, RRP44, MTR3, MTR4) and the RNase P (blue;

RRP14, RPP25, RPP38, RPP30, POP1). **f**, Dynamic profile of the human homologues of the yeast SSU processome proteins. **g**, Fold change of the large (blue) and small (green) ribosomal subunit proteins. **h**, Hierarchical clustering of 302 proteins using fold change data from Supplementary Table 2 (five and nine time point experiments). The indicated proteins are snRNP (SNRPA, SNRPD2, SNRPD3), Pol I (POLR1A, POLR1D, POLR1B, POLR1C) and exosome components. See Supplementary Fig. 2 for protein names. **i**, Comparison of fold change for a subset of proteins upon treatment with actinomycin D (red/green, average fold change after treatment for 80 and 160 min) and DRB (yellow, 80 min). **j**, Comparison of fold change for the small ribosomal subunit proteins upon treatment with actinomycin D (green) and MG132 (red, 8 h).

kinetics (Fig. 4g), consistent with previous reports that biogenesis and nuclear export of the large and small ribosomal subunits occur independently in budding yeast²¹. The variation in kinetics for different ribosomal proteins probably occurs because ribosome subunit assembly is a complex, multi-step process and not all the factors assemble on the rRNAs simultaneously. The dynamic profile of the human homologues of the yeast small-subunit (SSU) processome²², which is involved in the biogenesis of the 40S preribosomal subunit, displayed some of the most marked changes, decreasing by 10–15-fold in abundance within nucleoli after actinomycin D treatment (Fig. 4f). These data indicate that at least a subset of the processome components (homologues of Sof1, Upt11, Upt6, Upt7 and Upt14) are strongly dependent upon rRNA expression for their accumulation within nucleoli. We infer that these factors are specifically recruited into nucleoli to participate in ribosome subunit biogenesis and are not retained within nucleoli under conditions when no new rRNA substrate is being transcribed.

Figure 4h shows that changes in protein levels brought about by either recruitment to or loss from an organelle can be visualized and analysed by hierarchical clustering in a similar manner to the visualization of expression changes in microarray experiments²³. Inspection of the graph indicates reproducibility between patterns of the five and nine time point experiments. Groups of strongly recruited and depleted proteins clearly show the same pattern in both experiments (top and bottom clusters in Fig. 4h). Furthermore, functional and/or structural protein complexes such as the snRNPs, Pol I subunits and exosome cluster in tight regions (indicated in the rightmost column).

We repeated the kinetic analysis with different inhibitors that also affect nucleolar morphology. Treatment with 5,6-dichlorobenzimidazole riboside (DRB), which selectively inhibits RNA polymerase II but not polymerase I (ref. 24), resulted in nucleolar accumulation of a similar set of proteins to those seen in actinomycin D treatment (Fig. 4i and other data not shown). However, DRB caused fewer proteins to decrease in nucleolar abundance. Notably, the large and small ribosomal subunit proteins and components of the exosome remained relatively unchanged after DRB treatment. This supports the view that the decrease in ribosomal proteins and some other factors following inhibition of RNA polymerase I activity is caused, at least in part, by the export of ribosome subunits in the absence of new rRNA synthesis. We infer also that the accumulation of proteins in nucleoli following transcription inhibition results primarily from a block of RNA polymerase II activity.

Finally, we treated cells with the proteasome inhibitor MG132, which affects nucleolar morphology but does not directly inhibit RNA polymerase activity²⁵. A SILAC experiment was performed on HeLa^{YFP-p68} cells treated for up to 16 h with MG132 at a final concentration of 10 μ M. Quantitative proteomics showed that MG132 also caused a major change in the nucleolar proteome, but largely affected a different set of proteins as compared with inhibitors of transcription (Fig. 4j). MG132 caused a striking increase in the levels of ribosomal proteins in the isolated nucleoli, particularly in the case of small ribosomal subunit proteins, which contrasts with the general decrease in ribosomal proteins in nucleoli caused by actinomycin D. The reason for this unexpected effect is unclear but may reflect a novel regulatory link between ribosome biogenesis and protein degradation pathways, acting to balance rates of protein synthesis and breakdown.

The previously described effects of cellular growth and environmental conditions on nucleolar morphology and ribosome synthesis are reflected here by the large-scale and specific changes in the nucleolar proteome in response to distinct metabolic inhibitors. These data provide a more detailed and quantitative insight into how the cellular response to environmental stress and growth conditions affect the nucleolus. Further experiments will therefore be focused on the cause and effect of the relocalization of individual nucleolar proteins under these perturbations, and how

these redistributions can orchestrate the regulation of cell growth and proliferation.

This study establishes a quantitative approach for the high-throughput characterization of the flux of endogenous proteins through cellular organelles, demonstrated here for the nucleolus. It will be important to study the dynamic nature of other subcellular structures, compartments and organelles because it is likely that their protein compositions can also vary extensively under different growth and/or metabolic conditions. We conclude that there is no unique, complete proteome for the nucleolus, or probably for any other organelle, but rather an overlapping set of proteomes that are relevant to different cell states or conditions. $\square$

Methods

Isolation of stable isotope-labelled nucleolar proteins

Cells were grown for at least five cell divisions in L-arginine-, L-arginine-¹³C₆-¹⁴N₄-, or L-arginine-¹³C₆-¹⁵N₄-labelling media before drug treatment¹⁸. Actinomycin D (5 mg ml⁻¹ stock solution in EtOH) was added at a final concentration of 1 μ g ml⁻¹ to Arg6- and Arg10-labelled cells and incubated for 20 and 80 min, respectively. The experiment was repeated with drug treatment for 40 and 160 min of Arg6- and Arg10-labelled cells, respectively, to give a total of five time points with untreated Arg0 cells as a common zero time point. In independent experiments, cells were treated with actinomycin D to give a total of nine time points, and with the proteasome inhibitor MG132 (10 μ M for 1, 4, 8 and 16 h) and the polymerase II inhibitor DRB (25 μ g ml⁻¹ for 2 h and mock treated).

Nucleoli were isolated from HeLa cells or from a 1:1:1 mixture of Arg0, Arg6 and Arg10 HeLa^{YFP-p68} cells as previously described⁶ (<http://www.lamondlab.com/f5nucleolarprotocol.htm>). Isolated nucleolar proteins were separated on NuPAGE 4–12% Bis-Tris gel and excised into 16–20 slices. Peptides resulting from in-gel digestion were extracted from the gel pieces, desalted and concentrated on reverse-phase C18 tips, and eluted into 96-well plates for automated mass spectrometry analysis.

Western blots were performed with an antibody against p68 (gift from R. Janknecht), which recognized both the YFP-tagged and untagged forms of p68 (Fig. 4a) and an antibody against fibrillarin (gift from F. Fuller-Pace).

Mass spectrometry and data analysis

Mass spectrometric analysis was performed by liquid chromatography (Agilent HP1100) combined with tandem mass spectrometry (LC MS/MS) using a quadrupole time-of-flight instrument (QSTAR-XL, ABI-MDS-Sciex) or a linear ion-trap Fourier-transform ion-cyclotron resonance mass spectrometer (LTQ-FT-ICR, Thermo-Finnigan). For the QSTAR-XL, precursor ion spectra (m/z 350–1,500) and product ion spectra (m/z 70–1,500) of the four most intense ions were collected for 1 s. The LTQ-FT-ICR instrument was operated in the data-dependent mode to acquire high-resolution precursor ion spectra (m/z 300–1,500, $R = 25,000$ and ion accumulation to a target value of 10,000,000) in the ICR cell. The three most intense ions were sequentially isolated for accurate mass measurements by selected ion monitoring (SIM) scans (10 Da mass window, $R = 50,000$, and a target accumulation value of 50,000). The ions were simultaneously fragmented in the linear ion trap with a normalized collision energy setting of 27% and a target value of 2,000.

Stringent criteria were required for protein identification in the International Protein Index database using the Mascot program (Matrix Science) and LTQ-FT-ICR data: at least two matching peptides per protein, a mass accuracy within 3 p.p.m. (average absolute peptide mass accuracy was 0.7 p.p.m.), a Mascot score for individual peptides of better than 20, and a delta score of better than 5. Experiments with a reversed database²⁶ indicated that, under these conditions, proteins with two matching peptides were identified with a false positive rate of less than 0.1 per cent.

Protein ratios were calculated for each arginine-containing peptide as the peak area ratio of Arg6/Arg0 and Arg10/Arg0 of each single scan mass spectrum. The peptide ratios were averaged for all arginine-containing peptides sequenced for each protein and normalized to zero ($x - 1$). Normalized inverted ratios were calculated for ratios smaller than one [$1 - (1/x)$]. MS-Quant (<http://msquant.sourceforge.net/>), an in-house developed software program was used to evaluate the certainty in peptide identification and in peptide abundance ratio.

Live cell imaging

HeLa^{YFP-p68} cells⁸, HeLa YFP-FIB cells²⁷ and HeLa YFP-RPL27 cells²⁷ were cultured in Willco thin glass-bottomed microwell dishes (Intracel), mounted on a Deltavision Spectris microscope (Applied Precision) fitted in a transparent environmental chamber (Solent Scientific). Cells were imaged 60 $\times$ (NA 1.4) Plan Apochromat objective. Twelve optical sections separated by 0.5 μ m were recorded for each field and each exposure lasted for 0.05 s. After recording the first three time points, actinomycin D was added and cells were imaged for 2–3 h (SoftWoRx image processing software, Applied Precision). Nucleoli or nuclei (see Fig. 3) were outlined manually and five nucleoli/nuclei were measured from two independent experiments. In control experiments an equal volume of ethanol was added instead of actinomycin D.

Transcription assays

Briefly, isolated nucleoli were incubated with run-on buffer (100 mM KCl, 50 mM Tris-HCl pH 7.4, 5 mM MgCl₂, 0.5 mM EGTA, 0.5 mM ATP, 0.5 mM CTP, 0.5 mM GTP and 0.2 mM Br-UTP (Sigma Chemicals)) for 20 min at room temperature²⁸. The

incubated nucleoli were spotted on a poly-L-lysine coated slide (BDH) then fixed in 2% paraformaldehyde (5 min, room temperature) and permeabilized in 0.5% Triton X-100 (10 min, room temperature), before being immunolabelled with anti-BrdU antibody (Sigma Chemicals), FITC-conjugated anti-mouse immunoglobulin (Jackson's lab) and counterstained with Pyronin Y (Sigma Chemicals). Transcription in intact HeLa cells was detected as reported²⁹. Microscopy and image handling was performed using a Deltavision Spectris microscope as described above.

Received 18 October; accepted 16 November 2004; doi:10.1038/nature03207.

1. Visintin, R. & Amon, A. The nucleolus: the magician's hat for cell cycle tricks. *Curr. Opin. Cell Biol.* **12**, 752 (2000).
2. Guarente, L. Link between aging and the nucleolus. *Genes Dev.* **11**, 2449–2455 (1997).
3. Sherr, C. J. & Weber, J. D. The ARF/p53 pathway. *Curr. Opin. Genet. Dev.* **10**, 94–99 (2000).
4. Olson, M. O. Sensing cellular stress: another new function for the nucleolus? *Sci. STKE* [online] **pe10** (2004) (doi:10.1126/stke.2242004pe10).
5. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
6. Mauramatsu, M., Smetana, K. & Busch, H. Quantitative aspects of isolation of nucleoli of the Walker carcinosarcoma and liver of the rat. *Cancer Res.* **25**, 693–697 (1963).
7. Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
8. Andersen, J. S. *et al.* Directed proteomic analysis of the human nucleolus. *Curr. Biol.* **12**, 1–11 (2002).
9. Trinkle-Mulcahy, L., Sleeman, J. E. & Lamond, A. I. Dynamic targeting of protein phosphatase 1 within the nuclei of living mammalian cells. *J. Cell Sci.* **114**, 4219–4228 (2001).
10. Li, D., Meier, U. T., Dobrowolska, G. & Krebs, E. G. Specific interaction between casein kinase 2 and the nucleolar protein Nopp140. *J. Biol. Chem.* **272**, 3773–3779 (1997).
11. Yamamoto, R. T., Nogi, Y., Dodd, J. A. & Nomura, M. RRN3 gene of *Saccharomyces cerevisiae* encodes an essential RNA polymerase I transcription factor which interacts with the polymerase independently of DNA template. *EMBO J.* **15**, 3964–3973 (1996).
12. Charroux, B. *et al.* Gemin4. A novel component of the SMN complex that is found in both gems and nucleoli. *J. Cell Biol.* **148**, 1177–1186 (2000).
13. Lamond, A. I. & Earnshaw, W. C. Structure and function in the nucleus. *Science* **280**, 547–553 (1998).
14. Misteli, T. Protein dynamics: implications for nuclear architecture and gene expression. *Science* **291**, 843–847 (2001).
15. Lamond, A. I. & Sleeman, J. E. Nuclear substructure and dynamics. *Curr. Biol.* **13**, R825–R828 (2003).
16. Leung, A. K. & Lamond, A. I. *In vivo* analysis of NHPX reveals a novel nucleolar localization pathway involving a transient accumulation in splicing speckles. *J. Cell Biol.* **157**, 615–629 (2002).
17. Ong, S. E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).
18. Blagoev, B., Ong, S. E., Kratchmarova, I. & Mann, M. Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nature Biotechnol.* **22**, 1139–1145 (2004).
19. Perry, R. P. & Kelley, D. E. Inhibition of RNA synthesis by actinomycin D: characteristic dose-response of different RNA species. *J. Cell. Physiol.* **76**, 127–139 (1970).
20. Raghavan, A. *et al.* Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Res.* **30**, 5529–5538 (2002).
21. Tschochner, H. & Hurt, E. Pre-ribosomes on the road from the nucleolus to the cytoplasm. *Trends Cell Biol.* **13**, 255–263 (2003).
22. Dragon, F. *et al.* A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. *Nature* **417**, 967–970 (2002).
23. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
24. Tamm, I., Hand, R. & Caliguri, L. A. Action of dichlorobenzimidazole riboside on RNA synthesis in L-929 and HeLa cells. *J. Cell Biol.* **69**, 229–240 (1976).
25. Mattsson, K., Pokrovskaja, K., Kiss, C., Klein, G. & Szekely, L. Proteins associated with the promyelocytic leukemia gene product (PML)-containing nuclear body move to the nucleolus upon inhibition of proteasome-dependent protein degradation. *Proc. Natl Acad. Sci. USA* **98**, 1012–1017 (2001).
26. Olsen, J. V., Ong, S. E. & Mann, M. Trypsin cleaves exclusively C-terminal to Arginine and lysine residues. *Mol. Cell. Proteomics* **6**, 608–614 (2004).
27. Leung, A. K. *et al.* Quantitative kinetic analysis of nucleolar breakdown and reassembly during mitosis in live human cells. *J. Cell Biol.* **166**, 787–800 (2004).
28. Masson, C. *et al.* Conditions favoring RNA polymerase I transcription in permeabilized cells. *Exp. Cell Res.* **226**, 114–125 (1996).
29. Boisvert, F. M., Hendzel, M. J. & Bazett-Jones, D. P. Promyelocytic leukemia (PML) nuclear bodies are protein structures that do not accumulate RNA. *J. Cell Biol.* **148**, 283–292 (2000).
30. Scherl, A. *et al.* Functional proteomic analysis of human nucleolus. *Mol. Biol. Cell* **13**, 4100–4109 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. Fox for providing the HeLa^{YFP-p68} cell line and other members of the Lamond and the Mann laboratories for help and discussions. Work in the Center for Experimental Bioinformatics (CEBI) is supported by a grant from the Danish National Research Foundation. A.I.L. is a Wellcome Trust Principal Research Fellow and is funded by a Wellcome Trust Programme grant; A.K.L.L. was funded by a Croucher studentship; Y.W.L. was funded by The Human Frontier Science Program, which is also acknowledged for a network grant entitled 'Functional organization of the cell nucleus investigated through proteomics and molecular dynamics'.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.I.L. (a.i.lamond@dundee.ac.uk) or M.M. (mamm@bmb.sdu.dk).

Vitamin C degradation in plant cells via enzymatic hydrolysis of 4-O-oxalyl-L-threonate

Martha A. Green & Stephen C. Fry

The Edinburgh Cell Wall Group, Institute of Molecular Plant Sciences, School of Biological Sciences, The University of Edinburgh, Daniel Rutherford Building, The King's Buildings, Edinburgh EH9 3JH, UK

Increasing the L-ascorbate (vitamin C) content of crops could in principle involve promoting its biosynthesis or inhibiting its degradation. Recent progress has revealed biosynthetic pathways for ascorbate^{1–3}, but the degradative pathways remain unclear. The elucidation of such pathways could promote an understanding of the roles of ascorbate in plants⁴, and especially of the intriguing positive correlation between growth rate and ascorbate oxidase^{5,6} (or its products⁷). In some plants (Vitaceae), ascorbate is degraded via L-idonate to L-threarate (L-tartrate), with the latter arising from carbons 1–4 of ascorbate^{3,8–11}. In most plants, however (including Vitaceae)¹¹, ascorbate degradation can occur via dehydroascorbate, yielding oxalate¹² plus L-threonate, with the latter from carbons 3–6 of ascorbate^{3,10,13}. The metabolic steps between ascorbate and oxalate/L-threonate, and their subcellular location, were unknown. Here we show that this pathway operates extracellularly in cultured *Rosa* cells, proceeds via several novel intermediates including 4-O-oxalyl-L-threonate, and involves at least one new enzyme activity. The pathway can also operate non-enzymatically, potentially accounting for vitamin losses during cooking. Several steps in the pathway may generate peroxide; this may contribute to the role of ascorbate as a pro-oxidant^{14,15} that is potentially capable of loosening the plant cell wall and/or triggering an oxidative burst.

A significant proportion of a plant's ascorbate is found in the apoplast (the aqueous solution permeating the cell walls)^{6,16}. To investigate the degradation of apoplastic ascorbate, we fed 0.5 mM L-[1-¹⁴C]ascorbate to *Rosa* cell-suspension cultures. Extracellular ascorbate steadily disappeared, at a rate that was greatest in fast-growing cell cultures. In 5-day-old cultures, the initial rate of ascorbate degradation was ~10 $\mu\text{M min}^{-1}$. Over 90% of the ¹⁴C remained in the medium for the first 5 h, indicating extracellular metabolism rather than uptake by cells or loss as ¹⁴CO₂.

L-[1-¹⁴C]Ascorbate incubated with 5-day-old cultures yielded at least seven radioactive products (A–G), resolved by electrophoresis at pH 6.5 (see Fig. 4). The mobilities of these products relative to the marker Orange G (m_{OG} values) were: A, 0.00; B, 1.25; C, 1.38; D, 1.42; E, 1.96; F, 2.32; G, 3.15; and ascorbate, 1.01 (mean values from five electrophoretograms). m_{OG} is proportional to the ratio of a molecule's charge (Q) to its surface area (represented by $M_r^{2/3}$, ref. 17).

A and B were identified by comparison with markers as dehydroascorbate and 2,3-diketogulonate respectively; these are well-known ascorbate degradation products. On electrophoresis at pH 6.5 and 2.0, G co-migrates with oxalate. Its high mobility at pH 2.0 ($m_{\text{OG}} \approx 1.1$) is particularly characteristic of oxalate, whose first pK_a (≈ 1.25) is unusually low.

F is a novel compound. At pH 6.5, it migrates between L-threarate and erytharate (*meso*-tartrate), suggesting that F also has two negatively charged groups. At pH 2.0, its m_{OG} (0.83) is much higher than that for threarate (0.11) or erytharate (0.08), showing that F has a low pK_a .

Alkali rapidly hydrolysed ¹⁴C-F to [¹⁴C]oxalate (Fig. 1b) plus a non-radioactive product co-electrophoresing with L-threonate (Fig. 1a). High-pressure liquid chromatography (HPLC) showed

the hydrolysis product to be threonate, not erythrone (Fig. 1c, d), indicating that **F** is an *O*-oxalyl-threonate. **F** arose from L-ascorbate, so we conclude that it is an *O*-oxalyl-L-threonate (not the D-enantiomer).

To determine the position of its ester linkage, we periodate-oxidized ^{14}C -**F**. The radioactive products from 2-, 3- and 4-*O*- ^{14}C -oxalyl-L-threonates are predicted to have increased, unchanged and decreased m_{OG} values, respectively (according to their $Q/M_r^{2/3}$ ratios) (Fig. 2a). Periodate-treated ^{14}C -**F** gave one major ^{14}C -product, which had a lower m_{OG} than **F** (Fig. 2b), indicating that **F** is 4-*O*-oxalyl-L-threonate.

Products **C**, **D** and **E** were partially characterized. Non-volatile, radioactive products of [1- ^{14}C]ascorbate metabolism will probably contain two, three, four or six carbon atoms. Any five-carbon products formed by decarboxylation (for example, 2-keto-L-xylo-nate¹⁸) would be non-radioactive owing to loss of $^{14}\text{CO}_2$. At pH 6.5, **E** migrated slower than erythrone (a four-carbon compound with

two negative charges, that is, a $(\text{C}_4, 2-)$ compound) but faster than glyoxylate or glycolate, both $(\text{C}_2, 1-)$ compounds, suggesting that **E** is a $(\text{C}_6, 2-)$ compound. When **E** was eluted and re-electrophoresed under identical conditions, **C** and **E** were both found, indicating non-enzymatic conversion from **E** to **C** (Fig. 3, lanes d and f). Treatment with alkali converted the **C/E** mixture back to pure **E** (Fig. 3, lane e), suggesting that **E**, a $(\text{C}_6, 2-)$ compound, is a de-lactonised form of **C**, which is therefore a $(\text{C}_6, 1-)$ compound. At pH 6.5, **C** migrated only slightly faster than authentic 2-keto-D-gluconate, a $(\text{C}_6, 1-)$ compound, and slower than L-threonate, a $(\text{C}_4, 1-)$ compound, supporting the idea that **C** is a $(\text{C}_6, 1-)$ compound.

When a **C/D** mixture was eluted and re-run at pH 6.5, not only was **E** formed (by de-lactonization of **C**) but **F** and **G** were also found (Fig. 3, lanes a and c). Because a **C/E** mixture does not yield **F** or **G** (Fig. 3, lanes d–f), the **F** and **G** must arise from **D**, not **C**. Treatment with alkali completely converted the **C/D** mixture to **E**

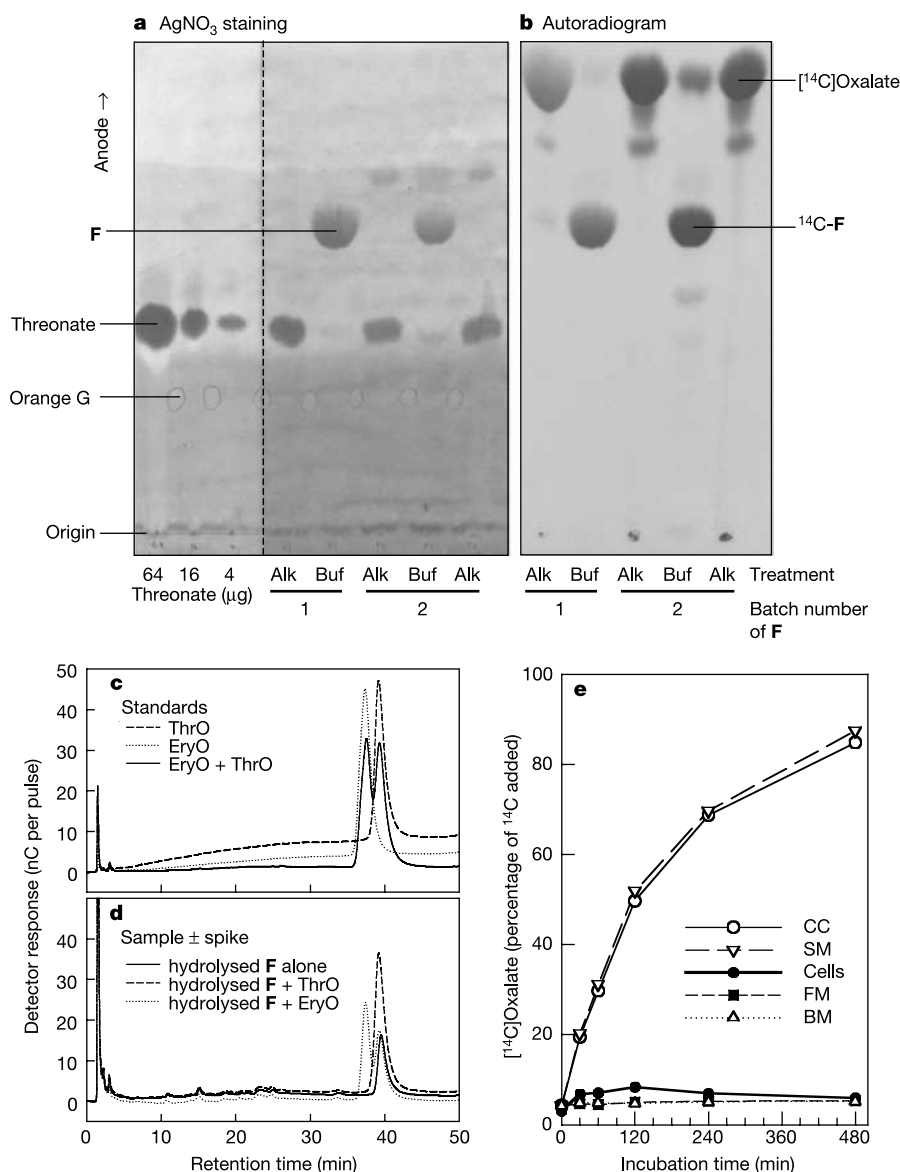


Figure 1 Chemical characterization and metabolism of product **F** (4-*O*-oxalyl-L-threonate). Low-specific-radioactivity ^{14}C -**F** (batches one and two) were treated with alkali (Alk) or buffer (Buf), electrophoresed at pH 6.5 (+EDTA), autoradiographed (**b**) and then stained with AgNO_3 (**a**). Authentic calcium L-threonate was also run. **c**, HPLC of authentic markers (ThrO, L-threonate; EryO, erythrone). **d**, HPLC of alkali-hydrolysed **F**, alone and

with 'spikes' of authentic L-threonate or erythrone. **e**, Fate of high-specific-radioactivity ^{14}C -**F** in freshly autoclaved medium (FM), whole *Rosa* cell-culture (CC), spent medium (SM) and boiled spent medium (BM). Also shown ('cells') is the ^{14}C that became associated with the cells in CC.

(from C) plus oxalate (from D and F) (Fig. 3, lane b).

F (C₆, 2-) arose from the much slower-migrating D, so we conclude that D is a (C₆, 1-) compound. To account for this, we suggest that spot D contained one or more cyclic oxalyl di-esters of L-threonate. The proposed reaction sequence is shown in Fig. 4e. To explain the reported *in vivo* incorporation of ¹⁸O from ¹⁸O₂ into threonate and from H₂¹⁸O into both threonate and oxalate¹⁰, we propose a reaction with O₂ (step 2) and two ester-hydrolysis reactions (steps 4 and 5).

To monitor the dynamics and enzyme-dependence of the pathway, we incubated 0.5 mM L-[1-¹⁴C]ascorbate with freshly autoclaved medium (Fig. 4a), *Rosa* cell-culture (Fig. 4b), cell-free spent medium (Fig. 4c) or boiled spent medium (Fig. 4d). This experiment was repeated three times; the results described below were obtained consistently.

The fate of [¹⁴C]ascorbate in freshly autoclaved medium (Fig. 4a) indicates that non-enzymatic reactions take place. Much of the [¹⁴C]ascorbate disappeared within 8 h and all seven products (A–G) were generated non-enzymatically. Product A was the first to be

formed; some A also arises artefactually during electrophoresis. The yield of ¹⁴C-A plateaued between 1 and 4 h. From 4 to 8 h, when most of the ascorbate had gone, the concentration of ¹⁴C-A declined (quantified from Fig. 4a), indicating dehydroascorbate turnover. ¹⁴C in B, F and G increased continuously from 0.5 to 8 h, whereas ¹⁴C in D plateaued after 0.5 h, suggesting D to F turnover. D was clearly resolved from C on electrophoretograms run for a longer time (not shown). Because the rate of ¹⁴C-B, ¹⁴C-F and ¹⁴C-G production remained high during the final 4 h interval, even though little ascorbate remained, we conclude that B, F and G were produced via dehydroascorbate, not directly from ascorbate. It has been shown that dehydro-L-[1-¹⁴C]ascorbate, but not 2,3-diketo-L-[1-¹⁴C]gulonate, yields [¹⁴C]oxalate *in vivo*¹². These observations support the pathway shown (Fig. 4e, steps 1–5 and 8–10).

Reactions taking place more rapidly in boiled spent medium compared with freshly autoclaved medium represent non-enzymatic steps that are promoted by heat-stable factors secreted by the cells (or that are inhibited by fresh-medium ingredients absorbed by the cells). [¹⁴C]Ascorbate disappearance in boiled spent medium equalled that in fresh medium, but the subsequent disappearance of [¹⁴C]dehydroascorbate and production of ¹⁴C-E, ¹⁴C-F and ¹⁴C-G were faster in boiled spent medium (Fig. 4d). The heat-stable factors affecting these steps remain unidentified.

Any reactions occurring more rapidly in spent medium (Fig. 4c) than in boiled spent medium suggest the presence of soluble extracellular enzymes. Such reactions included the initial disappearance of [¹⁴C]ascorbate, the production of ¹⁴C-G, and

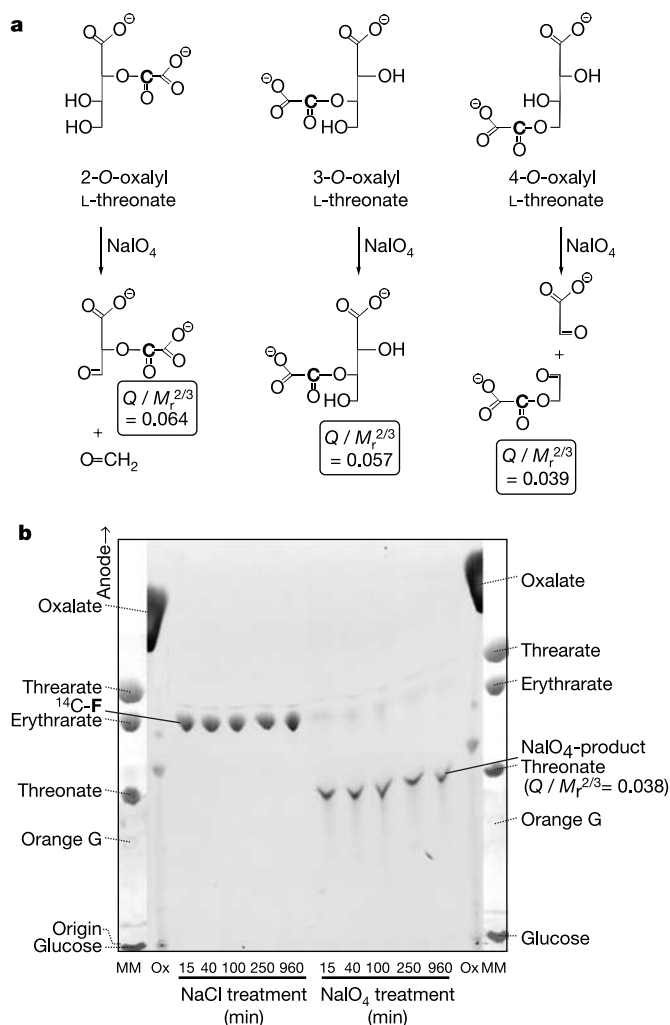


Figure 2 Periodate oxidation of ¹⁴C-F. **a**, Predicted reactions (if any) of three isomeric [¹⁴C]oxalyl-threonates with periodate. The $Q / M_r^{2/3}$ ratios are indicated for the expected radiolabelled products (assumed fully ionized). The position of the ¹⁴C atom (bold letter C) shown within the oxalyl group is arbitrary. **b**, Autoradiogram showing the effect of periodate on ¹⁴C-F. Aliquots of ¹⁴C-F were treated with NaCl or equimolar NaIO₄ for various periods, electrophoresed at pH 6.5 (+EDTA) and autoradiographed. Markers used: autoradiographed commercial [¹⁴C]oxalate (Ox) and a silver-stained marker mixture (MM).

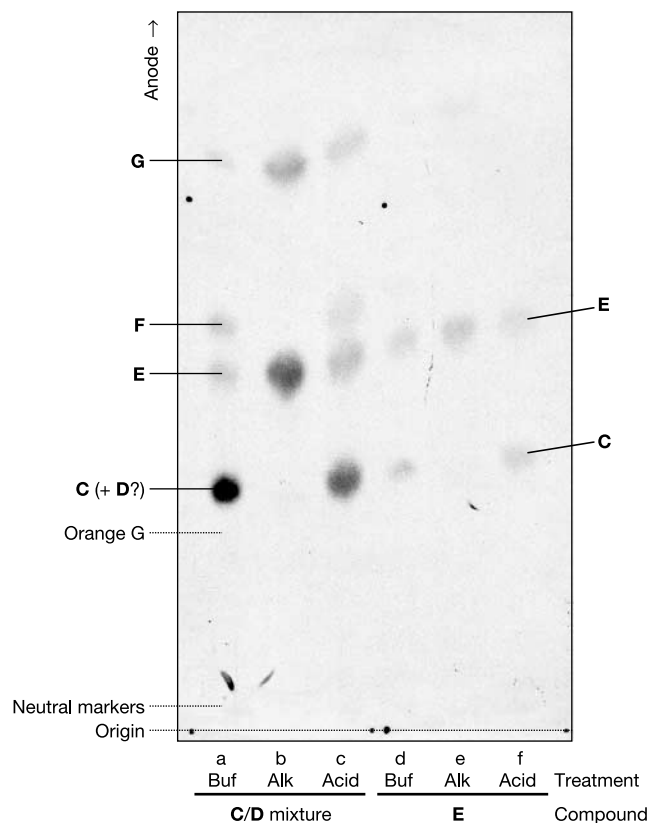
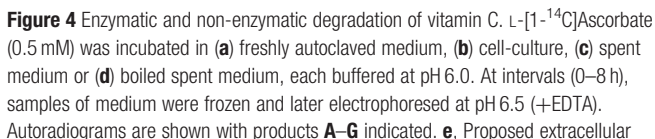


Figure 3 Inter-conversions involving compounds C, D, E, F and G (proposed chemical identities of D, F and G are shown in Fig. 4e). A mixture of C + D was eluted from a preparative electrophoretogram (pH 6.5), treated with buffer (pH 4.7), alkali or acetic acid, re-adjusted to pH 4.7 and finally re-electrophoresed at pH 6.5 (+EDTA). Pure compound E was similarly eluted, treated with buffer, alkali or acid, adjusted to pH 4.7 and re-electrophoresed. The autoradiogram is shown. The positions of Orange G, neutral markers and the origin (not visible) are indicated.

To monitor specifically the degradation of **F**, we incubated ^{14}C -**F** with cells or media. **F** was stable in freshly autoclaved medium and boiled spent medium, but rapidly hydrolysed in unboiled spent medium (Fig. 1e), indicating oxalyl esterase action (Fig. 4e, step 5). The presence of live cells (cell-culture) did not alter the results; therefore the esterase must act predominantly in solution in the

4-O-Oxalyl-L-threonate (F) is a new compound whose hydrolysis is catalysed by a soluble extracellular esterase. This activity may contribute to the biological role of apoplastic ascorbate. Although ascorbate is best known as an antioxidant, it can also be a pro-oxidant, generating reactive oxygen species including H_2O_2 (refs 14, 15). Its pro-oxidant activity may be increased by several steps in the proposed pathway (Fig. 4e). First, step 1 can be enzymatic⁶: $\text{AH}_2 + 1/2\text{O}_2 \rightarrow \text{A} + \text{H}_2\text{O}$ or non-enzymatic¹⁹: $\text{AH}_2 + \text{O}_2 \rightarrow \text{A} + \text{H}_2\text{O}_2$. Ascorbate loss in freshly autoclaved medium and boiled spent medium (Fig. 4a, d) indicates the non-enzymatic rate, whereas the total rate is shown in Fig. 4b, c. The non-enzymatic (H_2O_2 -generating) reaction contributed ~50% of the total rate. Second, step 2 in the pathway (Fig. 4e) may be partly oxidase-catalysed, and most oxidases generate H_2O_2 . Third, in tissues containing oxalate oxidase²⁰, step 6 of the pathway would yield H_2O_2 . Fourth and finally, in many plants, L-threonate yields L-threarate *in vivo* (Fig. 4e, step 7)¹³. Nothing is known of the enzyme(s) catalysing this four-electron oxidation. Theoretically, there could be two H_2O_2 -generating oxidases, with L-threo-tetruronate as intermediate. Although some evidence suggests that free L-threo-tetruronate is not an intermediate²¹, exogenous non-



pathway of ascorbate degradation. The radioactive carbon atom is shown in bold. Dashed arrows show reactions that are inconclusive or not detected in *Rosa* cultures. All steps except 6 and 7 occur at detectable rates non-enzymatically; some are accelerated by enzymes (see text).

radioactive L-threo-tetruronate did hinder the conversion of L-[¹⁴C]threonate to L-[¹⁴C]threarate²¹, suggesting that it may indeed participate.

Thus, one molecule of L-ascorbate potentially initiates a cascade of up to five molecules of apoplastic H₂O₂. Apoplastic H₂O₂, especially in the presence of some residual ascorbate, can generate hydroxyl radicals^{15,22}, which are proposed to loosen the cell wall, enhancing cell expansion^{23,24} and fruit softening²⁵. It is interesting that ascorbate secretion is promoted during the early stages of tomato fruit ripening²⁵.

Under other physiological circumstances, the pathway outlined in Fig. 4e might alternatively serve a protective, antioxidant role in which apoplastic ascorbate is sacrificially oxidized (for example, during ozone²⁶ or ultraviolet treatments²⁷). Some biological significance of compound F itself, for example as a signal, is also conceivable.

The pathway reported can operate non-enzymatically, so it cannot be either a taxonomic oddity of *Rosa* or a physiological oddity of cultured cells. The pathway's enzyme-independence suggests that it could operate not only in the apoplast (where we detected it), but also in other cellular compartments, perhaps being responsible for the synthesis of vacuolar oxalate crystals in idioblasts²⁸. □

Methods

Incubation of cells and media with [¹⁴C]ascorbate or [¹⁴C]-F

Rosa sp. ('Paul's Scarlet' rose) cell-suspension cultures, maintained as before²⁹, were subcultured fortnightly. Metal ions in fresh medium included 20 μM Fe³⁺, 4.5 μM Mn²⁺, 1.7 μM Zn²⁺, 0.08 μM Cu²⁺ and 0.04 μM Co²⁺. For radiolabelling experiments, 45 ml of 5-day-old culture was mixed with 5 ml 50 mM MES buffer (Na⁺), pH 6.0 (which is close to the natural pH of spent medium). The cell suspension was passed through muslin to remove large aggregates of cells, and the settled cell volume of the filtrate (cell-culture) was adjusted to 10% (v/v) by removal of some medium. Spent medium was obtained by filtration on fine nylon gauze to remove all cells; boiled spent medium was a sample of spent medium held at 100 °C for 5 min and then cooled. For this experiment, fresh medium was also supplemented with 5 mM MES, pH 6.0.

10 μl of aqueous L-[¹⁴C]ascorbate (Amersham; 0.52 MBq μmol⁻¹) was added to 143 μl of buffered medium (fresh, spent or boiled spent) or cell-culture in a round-bottomed glass tube (12-mm diameter), to give a final ascorbate concentration of 0.5 mM. The loosely capped tubes were shaken at 25 °C and 125 r.p.m. to maintain aeration. At intervals (0–8 h), a 20-μl aliquot of medium was added to 4 μl of 25 mM ascorbic acid (non-radioactive) to minimize further oxidation of [¹⁴C]ascorbate, and stored on liquid nitrogen until ready for electrophoresis. In the case of cell-culture, 20 μl of the cell suspension was filtered through a small plug of glass wool in a 200-μl pipette tip and the filtrate (~18 μl) was treated as for the other medium samples.

High specific-activity [¹⁴C]-F (~0.52 MBq μmol⁻¹) was eluted from electrophoretograms of the above samples, and fed back to cultures and medium as described for [¹⁴C]ascorbate but at a final concentration of 1.1 μM.

High-voltage electrophoresis

Samples were dried on Whatman 3MM paper and electrophoresed in buffer at pH 6.5 (1:33:300 by volume of acetic acid, pyridine and water, usually containing 5 mM EDTA), or at pH 2.0 (2:7:71 of formic acid, acetic acid and water), usually at 3.0–3.5 kV for 30 min. More prolonged electrophoresis was used to separate compounds C and D. The papers were cooled with toluene (for pH 6.5) or white spirit (for pH 2.0) during the run³⁰. Orange G (5 μg) was used as an internal reference marker, and electrophoretic mobilities are reported as *m*_{OG} values (that is, mobility corrected for electro-endo-osmosis, relative to that of Orange G). Colourless markers were stained with AgNO₃ or bromothymol blue³⁰. Autoradiography was on Kodak BioMax MR-1 and the radioactive spots were later quantified by scintillation counting. Authentic markers were commercial except dehydroascorbate (prepared by treatment of ascorbate with ascorbate oxidase) and 2,3-diketogulonate (prepared by treatment of dehydroascorbate with 100 mM NaOH for 5 min at 25 °C, neutralized with acetic acid and used immediately). Radioactive samples required for further analysis were eluted from the electrophoretogram with water after removal of any scintillant by washing in toluene and drying³⁰.

Alkali treatment of novel compounds

A sample of low-specific activity [¹⁴C]-F (batch one), eluted from a preparative electrophoretogram, was dispensed as 97-Bq (~50-μg) aliquots, which were dried and treated with either (1) 10 μl of 100 mM NaOH for 40 min at 25 °C, to cause alkaline hydrolysis, followed by 12 μl 200 mM acetic acid, or (2) 12 μl 200 mM acetic acid pre-mixed with 10 μl 100 mM NaOH (buffer pH ≈ 4.5). For [¹⁴C]-F (batch two), each aliquot was 590 Bq (~30 μg); otherwise, treatments were identical. Products were electrophoresed at pH 6.5 (+EDTA). A duplicate sample of NaOH-treated F (batch one) was analysed by HPLC on Dionex CarboPac PA1 with the following eluent profile of NaOH concentrations at 1 ml min⁻¹: 0–5 min, 10 mM; 5–25 min, 10–40 mM; 25–45 min, 40 mM; 45–50 min,

40–80 mM; 50–60 min, 80 mM; 60–70 min, 10 mM. A Dionex pulsed amperometric detector with gold electrode was used.

Samples of E or C + D were treated with 100 mM NaOH at 25 °C for 16 h then adjusted to pH 4.7 with acetic acid and immediately analysed by electrophoresis at pH 6.5. Control samples were given the same quantities of acid and NaOH but in reverse order (or simultaneously) so that no alkaline hydrolysis occurred.

Periodate treatment of compound F

[¹⁴C]-F (batch one; 50 μg) was dissolved in 100 μl ice-cold 250 mM formate (NH₄⁺) buffer (pH 3.7) containing 50 mM NaIO₄, and incubated in the dark at 0 °C (ref. 30). At intervals, 20 μl was added to 2 μl ethane-1,2-diol to destroy remaining periodate. A second 50-μg sample received 50 mM NaCl instead of NaIO₄. Products were electrophoresed at pH 6.5 (+EDTA).

Received 20 August; accepted 4 November 2004; doi:10.1038/nature03172.

Published online 19 December 2004.

- Wheeler, G. L., Jones, M. A. & Smirnoff, N. The biosynthetic pathway of vitamin C in higher plants. *Nature* **393**, 365–369 (1998).
- Agius, F. et al. Engineering increased vitamin C levels in plants by overexpression of a D-galacturonic acid reductase. *Nature Biotechnol.* **21**, 177–181 (2003).
- Asard, H., May, J. M. & Smirnoff, N. (eds) *Vitamin C Function and Biochemistry in Animals and Plants* (Bios Scientific Publishers, London, 2004).
- Smirnoff, N. The function and metabolism of ascorbic acid in plants. *Ann. Bot.* **78**, 661–669 (1996).
- Lin, L. S. & Varner, J. E. Expression of ascorbic acid oxidase in zucchini squash (*Cucurbita pepo* L.). *Plant Physiol.* **96**, 159–165 (1991).
- Pignocchi, C., Fletcher, J. M., Wilkinson, J. E., Barnes, J. D. & Foyer, C. H. The function of ascorbate oxidase in tobacco. *Plant Physiol.* **132**, 1631–1641 (2003).
- Hidalgo, A., García-Herdugo, G., González-Reyes, J. A., Morré, D. J. & Navas, P. Ascorbate free-radical stimulates onion root growth by increasing cell elongation. *Bot. Gaz.* **152**, 282–288 (1991).
- Williams, M. & Loewus, F. A. Biosynthesis of (+)-tartaric acid from L-ascorbic-4-C¹⁴ acid in grape and geranium. *Plant Physiol.* **61**, 672–674 (1978).
- Saito, K. & Kasai, Z. Synthesis of L-(+)-tartaric acid from L-ascorbic-acid via 5-keto-D-gluconic acid in grapes. *Plant Physiol.* **76**, 170–174 (1984).
- Saito, K., Ohmoto, J. & Kuriha, N. Incorporation of ¹⁸O into oxalic, L-threonic and L-tartaric acids during cleavage of L-ascorbic and 5-keto-D-gluconic acids in plants. *Phytochemistry* **44**, 805–809 (1997).
- deBolt, S., Hardie, J., Tyerman, S. & Ford, C. M. Composition and synthesis of raphide crystals and druse crystals in berries of *Vitis vinifera* L. cv. Cabernet Sauvignon: ascorbic acid as precursor for both oxalic and tartaric acids as revealed by radiolabelling studies. *Aust. J. Grape Wine Res.* **10**, 134–142 (2004).
- Yang, J. C. & Loewus, F. A. Metabolic conversion of L-ascorbic-acid to oxalic-acid in oxalate-accumulating plants. *Plant Physiol.* **56**, 283–285 (1975).
- Helsper, J. P. & Loewus, F. A. Metabolism of L-threonic acid in *Rumex* × *acutus* L. and *Pelargonium crispum* (L.) l'Hér. *Plant Physiol.* **69**, 1365–1368 (1982).
- Podmore, I. D. et al. Vitamin C exhibits pro-oxidant properties. *Nature* **392**, 559 (1998).
- Fry, S. C. Oxidative scission of plant cell wall polysaccharides by ascorbate-induced hydroxyl radicals. *Biochem. J.* **332**, 507–515 (1998).
- Takahama, U. Redox state of ascorbic acid in the apoplast of stems of *Kalanchoe daigremontiana*. *Plant Physiol.* **89**, 791–798 (1993).
- Offord, R. E. Electrophoretic mobilities of peptides on paper and their use in the determination of amide groups. *Nature* **211**, 591–593 (1966).
- Deutsch, J. C. Oxygen-accepting antioxidants which arise during ascorbate oxidation. *Anal. Biochem.* **265**, 238–245 (1998).
- Dekker, A. O. & Dickinson, R. G. Oxidation of ascorbic acid by oxygen with cupric ion as catalyst. *J. Am. Chem. Soc.* **62**, 2165–2171 (1940).
- Lane, B. G., Dunwell, J. M., Ray, J. A., Schmitt, M. R. & Cuming, A. C. Germin, a protein marker of early plant development, is an oxalate oxidase. *J. Biol. Chem.* **268**, 12239–12242 (1993).
- Saito, K. Metabolism of L-threotetruronic acid by *Pelargonium crispum*. *Phytochemistry* **31**, 1219–1222 (1992).
- Halliwell, B. & Gutteridge, J. M. C. Role of free radicals and catalytic metal ions in human disease: an overview. *Methods Enzymol.* **186**, 1–85 (1990).
- Schopfer, P. Hydroxyl radical-induced cell-wall loosening *in vitro* and *in vivo*: implications for the control of elongation growth. *Plant J.* **28**, 679–688 (2001).
- Rodríguez, A. A., Grunberg, K. A. & Taleisnik, E. L. Reactive oxygen species in the elongation zone of maize leaves are necessary for leaf extension. *Plant Physiol.* **129**, 1627–1632 (2002).
- Dumville, J. C. & Fry, S. C. Solubilisation of tomato fruit pectins by ascorbate: a possible non-enzymic mechanism of fruit softening. *Planta* **217**, 951–961 (2003).
- Ploch, M., Lyons, T., Ollershaw, J. & Barnes, J. Simulating ozone detoxification in the leaf apoplast through the direct reaction with ascorbate. *Planta* **210**, 454–467 (2000).
- Conklin, P. L., Williams, E. H. & Last, R. L. Environmental stress sensitivity of an ascorbic acid-deficient *Arabidopsis* mutant. *Proc. Natl Acad. Sci. USA* **93**, 9970–9974 (1996).
- Keates, S. E., Taryl, N. M., Loewus, F. A. & Franceschi, V. R. L-Ascorbic acid and L-galactose are sources for oxalic acid and calcium oxalate in *Pistia stratiotes*. *Phytochemistry* **53**, 433–440 (2000).
- Fry, S. C. & Street, H. E. Gibberellin-sensitive suspension cultures. *Plant Physiol.* **65**, 472–477 (1980).
- Fry, S. C. *The Growing Plant Cell Wall: Chemical and Metabolic Analysis* Reprint edn (Blackburn Press, Caldwell, New Jersey, 2000).

Acknowledgements We thank B. Dudley and J. Miller for technical assistance. M.A.G. thanks the BBSRC for a research studentship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.C.F. (S.Fry@Ed.Ac.UK).

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayouni (4978)

UK/ RoW/ Ireland:

Andy Douglas (4975)
Nils Moeller (4953)
Irene Vigila-Atton (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Sille Opstrup (4994)

France/ Switzerland:

Amelie Pequinot (4974)

Advertising Production

Manager: Billie Franklin
To send materials use London
address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Niamh Shields

European Satellite Office

**Germany/ Austria/ Italy/
The Netherlands/ Belgium:**

Patrick Phelan
Tel + 49 89 54 90 57 11
Fax + 49 89 54 90 57 20
e-mail: p.phelan@nature.com
Sharon de Weert
Tel + 44 (0) 20 7843 4970
e-mail: s.deveert@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatomachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746
Asia-Pacific Sales Director:
Rinoko Asami
e-mail: rasami@naturejpn.com

naturejobs

Responses and resolutions

At the end of last year, in between trips, interviews and deadlines, I checked the *Naturejobs* feedback inbox and was dismayed to find it clogged with spam. After deleting hundreds, if not thousands, of offers for dubious goods and services, I resolved not to let this happen again in 2005 — and to address the four most pertinent questions the box contained.

One reader said it was irresponsible to recommend that young scientists pursue higher-risk research, rather than following the crowd. I'd respond by saying that every course of action has its risks and one needs to calculate them and match them with one's goals. Another reader had a question about the wisdom of using Francis Crick as an example of someone who changed scientific direction at a late age and was successful. Again, my main point was that such a move is possible; my only caveat is to weigh the risks of making such a move before doing it.

Another column, about US President Bush's impact on foreign scientists, had one writer seeing red, and accusing me of being partisan. She said that it was unfair to attribute the decline in numbers of foreign graduate students in the United States to politics. Although I agree that correlation does not equal cause-and-effect, I'd say that the decline — happening after the president changed his visa policy — is impossible to ignore.

My favourite letter, though, is from someone who said that the four graduate students who kept journals for *Naturejobs* last year all sounded too positive. I'd respond by saying that they did report bumps in the road, but carried out a good job in writing about how this would affect their future career decisions — the entire point of the column. And I'd add that, in addition to resolving to address readers' concerns faster and more regularly this year, *Naturejobs* will attempt to offer viable solutions to scientific career issues and not just complain about problems.

Paul Smaglik
Naturejobs editor



Contents

CAREER VIEW

Nuts & Bolts
Online applications
Graduate Journal
Brave old world
Movers
Richard Perham

p90

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

GRADUATE JOURNAL

Brave old world

Growing up in the Thuringian Forest in East Germany, I faced the Berlin Wall from the wrong side until it collapsed in 1989. Eight years later, I headed west to study medicine in Würzburg, Bavaria. At the time, western Germany seemed significantly different from the east: there was less urban decay, for instance.

But the differences between eastern and western Germany were minor compared with what I've just experienced trading Würzburg for Oxford. After I graduated last year, I decided to pursue neuroscience research rather than become a clinician. A Wellcome Trust studentship helped me take the next step, and brought me to Britain.

Arriving in Oxford was like entering another dimension — the place seems more like an ancient Greek city-state. It has its own traditions, myths, laws and anachronisms. Imagine being tested on the latest developments in molecular biology while wearing an academic gown, dark suit, white bow tie and plain white shirt.

Cutting-edge science and traditional dress, innovative research and rigid traditions. That's how Oxford is — a living contradiction. It's one of the most fascinating places I've been. Life here spans several centuries and lifestyles. It makes Oxford a special place. And it makes moving here the biggest change in my life so far. ■

Tobias Langenhan is a first-year graduate student in neuroscience at the University of Oxford, UK.

NUTS & BOLTS

Online applications

Ever wonder if anyone actually reads the CVs that you submit online in response to a job advert? Because many companies have switched to electronic systems for tracking applications, it is important to use their established framework. But how do you ensure that your bid for employment gets noticed?

Before clicking on 'submit', make sure you have customized your application for the position. This may seem like a lot of work, but it's time well spent. First the job-tracking software and later the hiring manager will be scanning for as exact a match as possible.

Does this mean you should inflate or fabricate to fit? No. Be true to yourself and show your best side by highlighting the actions and results from your experience that are most likely to meet the employer's needs. Link the



With Deb Koen
Careers consultant

words you use to the terminology in the job listing — this will help you make the initial cut. And tailor your message by using the space that's generally provided for a cover letter to focus your comments on the potential value that you offer the position and employer.

Beyond customizing your materials, you must go 'live' at every opportunity and as quickly as possible. Although the front end of the application process is immersed in technology, the hiring decisions are always made by humans. Go out of your way to make

contacts that may be helpful to you to learn about organizations, to generate leads and to make referrals. Employers love it when someone they respect makes a referral to them because it simplifies the selection process.

Approach your own network of fellow researchers, former advisers and colleagues for possible referrals that may draw attention to your online application. And be sure to return the favour. Networking is an ongoing, mutually beneficial exchange that builds lasting relationships (see *Nature* 430, 812–813; 2004).

With hundreds, sometimes thousands, of CVs submitted online, customizing your message and making contact can turn a virtual application into a concrete opportunity. ■

Deb Koen is vice-president of Career Development Services and a columnist for *The Wall Street Journal's CareerJournal.com*.

MOVERS Richard Perham, master of St John's College, University of Cambridge, UK



The new master of St John's College looks out of his window onto the ancient Cambridge gardens and says: "To this day, I think I'm lucky to be here." It's more than just a modest remark from Richard Perham, former head of the University of Cambridge biochemistry department.

Brought up by a widowed mother, Perham became the first person in his family to go to university, with the help of scholarships. Although he has enjoyed stints in Australia, mainland Europe and the United States (he met his wife, the biologist Nancy Lane, at Yale), he's remained based at Cambridge since he arrived there fresh from national service in

1958. He has always found inspiration, he says, "walking down these streets that Newton and Darwin walked down, in this place where so much has happened".

Robust funding at that time created an innovative environment. "With money around, one could do all sorts of things," says Perham, recalling the start of his work on self-assembling structures, multi-functional proteins and macromolecular complexes. "It was a more rosy future than young people perceive now."

Key interactions with legendary scientists further fuelled Perham's enthusiasm. Max Perutz was an early mentor. And Fred Sanger — on his way to a second Nobel chemistry prize — taught Perham to choose a problem that didn't have an obvious solution.

"It was a lively place, a lot of fun," Perham remembers. "At the same time, I was impressed by people's dedication to research." He has tried to create that same atmosphere in his own labs, and

draws on memories of the generosity of Perutz and Sanger when he helps his own students and postdocs.

"I always remind them that they work with me, not for me," he says. "They'll make their own mistakes, and that's part of the learning process."

It's crucial, Perham adds, for group leaders to learn from their students too. "One goes to a lot of trouble to recruit the brightest and best, and it would be silly not to recognize that they have very acute minds," he says.

Perham is still involved in research, collaborating with colleagues at the University of California, Irvine, to make safer forms of vaccine that contain none of the original pathogen. But his new concern as master of St John's is to raise funds for scholarships and bursaries. If getting his chance at Cambridge was "lucky", he is determined to spread that luck around as many bright young hopefuls as possible. ■

CV **1964–2004:** University of Cambridge biochemistry department (rising to head of department in 1985)
1961–1964: PhD, Medical Research Council Laboratory of Molecular Biology, Cambridge, UK